

Uncertainty and naivety in Germany's science politics

The SPD and Green coalition that governs Germany has made notable progress and demonstrated flexibility in the process. But the Greens' lack of expertise needs redressing if good science is to get the support it deserves.

An agreement last week within the German coalition government to ban the reprocessing of nuclear waste is a small milestone in German politics — a goal to which the majority of Social Democrats were always committed, but whose attainment was accelerated by their junior coalition partners, the Green party. Should researchers be worried about the impact of Green politics on other technologies and on science?

So far the signs are encouraging. Now nearly 100 days into their first government, the Greens have already shown themselves willing to compromise to stay in power, for example over 'eco-taxes' on power stations and on exempting reactors producing neutrons for research from measures against nuclear energy (see page 189). But they have made clear that their fight against the FRM II research reactor, arrogantly and ill-advisedly designed by the Technical University of Munich to burn weapons-grade, highly enriched uranium, is not over.

German scientists are now nervously waiting to see how the Greens and the Social Democrats will reach a compromise on another sensitive issue: agricultural biotechnology. The Greens tolerate, and the Social Democrats celebrate, biomedical science and its applications, now flourishing in Germany, thanks particularly to former research minister Jürgen Rüttgers.

But the equivocal and hostile attitudes expressed respectively by the Social Democrats and the Greens towards agricultural biotechnology leave plant scientists in a state of uncertainty. One issue is a promised review of how the licensing of field trials of transgenic plants

and marketing of transgenic foods are organized. The Greens want this responsibility to be transferred from the Robert Koch Institute, an office of the federal health ministry, to the federal environment ministry. Plant scientists worry that any change in the system, however logical it might seem, could allow the Greens to introduce new restrictions, or at least place licensing in inexperienced hands. The field-trial licensing office in the Robert Koch Institute is rightly respected for its expertise and reliability. At the same time, the sophistication of the Green party in this area is questionable. The change should be resisted.

The second issue is the fate of the newly created plant genome programme GABI, which was launched by Rüttgers last autumn as an attempt to do for agricultural biotechnology what he has already done for medical biotechnology. The scale of the programme has never been defined, and researchers now fear that it could be severely restricted. Serious money would be needed to make GABI competitive with activities in other countries.

The root cause for concern is the fact that the parliamentary Green party has no serious expertise in biotechnology, having ensured that its only expert, former parliamentarian Manuel Kiper, was shunted well down the party list before last September's elections. Since the party has a strong, and negative, political stance on agricultural applications of biotechnology, it needs to fill this regrettable gap as fast as possible. It must recognize that GABI will deliver important basic research. The Greens' ability to compromise is welcome, but given its influence, its scientific and technical naivety is unacceptable. □

Arima ascendant

The choice of the new head of Japan's education ministry and Science and Technology Agency is to be applauded.

The decision of Japan's prime minister Keizo Obuchi to retain Akito Arima, a physicist and former president of Tokyo University, as minister of education and also appoint him as head of the Science and Technology Agency in his new coalition government (see page 188) is a wise move that should help smooth the complex impending merger of the agency and the ministry.

Arima, a champion of reform of Japan's public-sector research system with experience of working with both the education ministry and the agency in his former roles as president of Tokyo University and head of the agency's Institute of Physical and Chemical Research (RIKEN), is ideally qualified for the job. He is also rare among Japan's scientists in having the political savvy and determination to carry out his difficult task.

The merger of the education ministry and agency scheduled for 2001 will not be easy. The organizations are as alike as oil and water and have often clashed in the past. The education ministry is large, conservative and devoted primarily to issues of education in schools and universities rather than scientific research. The much smaller and younger Science and Technology Agency is a comparatively lightweight organization in the political world that has devoted most of its

resources to large-scale science and technology projects in space, marine science and nuclear energy.

One area where rapid progress can be expected under Arima will be the introduction of a much-needed nationwide system for assessing research in universities, as well in institutes of the agency. Arima has been a pioneer of such assessment in his former roles at Tokyo University and RIKEN. He is also a strong advocate of building greater public awareness and understanding of science and of more creative science curricula in schools. All of which is heartily to be welcomed.

A delicate task for Arima will be rationalization of overlapping functions of the agency and ministry, particularly in the areas of marine and space science. While a case might be made for greater sharing of resources between the agency's Japan Marine Science and Technology Center and the Ocean Research Institute of Tokyo University, there is greater need for caution with space science, where the internationally renowned Institute of Space and Astronautical Science under the ministry of education could be seriously impaired by any attempt to merge it with the agency's huge National Space Development Agency, which is run by engineers. Arima must retain the best interests of science in dealing with these complex issues. □



Embryonic stem-cell research exempt from ban, NIH is told

[WASHINGTON] The US Department of Health and Human Services (DHHS) has issued a legal opinion saying that research on human embryonic stem cells does not fall under the ban on federal funding for human embryo research. The department says this is because such cells do not constitute an 'organism' as described in the legislation.

The decision was due to be announced by Harold Varmus, the director of the National Institutes of Health (NIH), at a meeting of the National Bioethics Advisory Commission on Tuesday (19 January). It marks the next move in a delicate political course the NIH is trying to steer in the face of an uncertain degree of conservative opposition.

The decision could be codified by a bill likely to be introduced this week by Senator Arlen Specter (Republican, Pennsylvania), chair of the appropriations subcommittee that funds NIH. The bill allows federal support for scientists doing research on stem cells, and may go further, allowing funding for those who derive stem cells from embryos.

Last November, two teams of scientists announced the isolation of the cells, generating hopes for their use in cell and tissue transplants, drug development and basic developmental biology (see *Nature* **396**, 104; 1998). But the uncertain moral status of such cells, given their close link to human embryos, led opponents of human embryo research to urge that federal dollars should not be used to support their use in research.

Bolstered by the DHHS decision, the NIH plans to draw up guidelines for investigators applying for NIH money to do stem-cell research, and to form an oversight group to ensure that applicants follow the guidelines.

"This opinion allows us to proceed carefully and thoughtfully with a line of research that has enormous potential for the treatment of almost every disease and condition," says an NIH official.

But the official added that NIH-funded scientists should wait until the guidelines are in place before launching human stem-cell research. NIH officials hope that, by demonstrating careful guidance of the stem-cell field, they will win public respect for it. They had no such opportunity with human embryo research because of the funding ban.

Varmus "knows the public wants to hear" about the work, says the NIH official, and plans to report publicly on how many researchers are being funded with how many dollars, as well as describing the nature of their work and their findings.

The DHHS decision interprets an existing law that bars federal funding for human



Varmus: planning guidelines for grant applicants.

embryo research in a way that allows support for research on stem cells, but not for their extraction from human embryos. Varmus received the legal opinion from Harriet Rabb, the general counsel at the DHHS, of which NIH is a part.

Rabb noted that the ban on government funding for embryo research describes a human embryo as an "organism" — defined scientifically as an individual constituted to carry out all the functions of its species. Rabb decided that, because stem cells are not organisms, they are not covered by the ban.

The NIH explained the rationale for the decision in the context of the work of James Thomson of the University of Wisconsin, who derived embryonic stem-cell lines using blastocysts left over from fertility treatments.

Thomson derived his cells from the inner cell mass of the blastocyst, which consists of cells that, although retaining the ability to differentiate into many kinds of cells, are unable to give rise to an embryo if implanted in a uterus. So, according to a paraphrase of the DHHS decision provided by NIH, "even if the cells are derived from a human embryo, they are not themselves a human embryo".

Rabb advised, however, that work to derive the cells does fall under the ban, which has been attached by Congress to the annual NIH spending bills since 1995. She confirmed that the existing law allows, within certain constraints, federal funding for deriving stem cells from tissue of aborted fetuses.

This achievement by John Gearhart, a developmental geneticist at Johns Hopkins University School of Medicine in Baltimore, Maryland, was also announced in November. Gearhart and Thomson were funded by

Insider trading alert over bioscience companies

[SAN FRANCISCO] Officials of the US Securities and Exchange Commission (SEC) have begun to look closely at bioscience stocks for illegal trading.

Over the past two years, the agency has filed six cases against scientists and consultants suspected of having used inside knowledge about research results to trade stock or to suggest investments to family and friends.

"The agency is increasingly focused on biotechnology as an area where insider trading is a severe problem," says James R. Ferguson, a Chicago securities attorney.

Ferguson says the scientific practice of broadly disseminating confidential research results that could be important to a company's future helps to create the potential for misbehaviour.

Peer reviewers, researchers in clinical trials, journalists and company consultants often have early access to sensitive data.

Such a situation is especially dangerous in young biosciences companies because a positive or negative clinical trial result can have a huge impact on the fate of a company's stock, says Ferguson. And, because scientific developments in the field are more widely reported in the popular press than in other technology areas, they have a greater effect on the stock price.

Questions have been raised, for example, by the frenetic trading in Geron Corp. the day before a report in *Science* last November that company-funded researchers had been able to grow human embryonic stem cells successfully in the

laboratory (see above).

On the day on which the article appeared, the value of Geron stock nearly tripled, although such timing can be purely coincidental. (It is SEC policy to decline to comment on whether investigators are reviewing trading activity in a particular company.)

One industry analyst says that insider trading rules can be especially tricky in the biosciences. "The majority of biotechnology companies don't have revenues or profits, so [investment decisions] revolve around deal rumours, or what you can find out about what's going on experimentally with these drugs," he says.

Some companies, recognizing the potential for problems, issue a summary of clinical results as soon as possible, even before publication in a peer-reviewed journal.

Sally Lehrman

Geron, a biotechnology company in Menlo Park, California.

An early draft of Specter's bill that was circulating last week declares that "the Secretary [of Health and Human Services] may conduct, support or fund research on, or utilizing, human embryonic stem cells". It also says that federal funding could support the derivation of embryonic stem cells from leftover embryos resulting from *in vitro* fertilization, providing consent was given by the couple. But it was not clear that this provision would remain in the bill.

Specter indicated at a Senate hearing last week that he intended to try to lift the human embryo research ban as it applies to stem-cell research "at a very early stage" because of the potential for it to deal with serious diseases.

At the hearing, witnesses who implored Specter to exempt stem-cell research from the federal ban included a young man diagnosed with Parkinson's disease at the age of 27 and Doug Melton, chairman of the department of molecular and cellular biology at Harvard University, who has a seven-year-old son with juvenile diabetes. Both

diseases are among those for which stem-cell research is thought to hold most promise.

The biomedical community applauded the DHHS decision. "We are delighted," says William Brinkley, president of the Federation of American Societies for Experimental Biology. "This makes it possible for many more investigators in this country to have access to this technology." Ron Eastman, chief executive officer of Geron, calls it "good news for science and medicine".

But Richard Doerflinger, a spokesman for the National Conference of Catholic Bishops, protested that the decision means that the government will be providing incentives for embryo destruction. "The reward for destroying them is an NIH grant to work on the stem cells thus produced," he said. "It doesn't matter what you did to obtain the stem cells as long as whatever destruction is needed was done without federal funds."

Doerflinger noted that the law governing the use of fetal tissue in federally funded research prohibits carrying out abortions in order to get the tissue. A woman must have chosen an abortion for unrelated reasons,



Specter: bill would allow funding for research on human embryonic stem cells.

and have no contact with the researcher. He says that destroying an embryo to obtain stem cells is morally equivalent to an abortion, and the new policy therefore contravenes the spirit of the existing law on the use of fetal tissue.

Conservative Republicans in Congress could challenge the DHHS decision by broadening the existing ban to explicitly include stem-cell research. **Meredith Wadman**

Space station faces research cuts to cover risk of Russian default

[WASHINGTON] Space-station managers at the US space agency NASA are braced for a 30 per cent reduction in funds next year for research on the station. The lower-than-expected allocation is to enable the agency to guard against a possible Russian default on delivering key elements of the station.

Although NASA's budget for the fiscal year 2000 will not be finalized until 1 February, the agency has been told by the White House Office of Management and Budget (OMB) to expect less than half the money it requested to cover such a situation.

OMB's refusal of the full request for contingency planning would result in a "considerable reduction in funding to the [space station] research programmes," says Michael Suffredini, manager of the space-station payloads office at NASA's Johnson Space Center in Houston.

A memorandum sent by Suffredini to agency research managers on 8 January was leaked last week to *NASA Watch*, an independent website that tracks space policy issues. As a result of OMB's constraints, the space-station programme faces a \$200 million shortfall over the next five years, plus an additional burden of \$70 million to pay for the development of new technology.

The easiest place to find the money without jeopardizing the space station's tight construction schedule is the research 'utilization' budget, which pays for NASA-funded scientists to develop hardware and experiments for the station. Agency science managers have therefore been asked to draw up plans to scale down their research

programmes.

Top priority will be given to maintaining the launch schedule for large 'facility class' instruments, which will be used by many different researchers on the station. Priority will also be given to building hardware for individual experiments that have already passed key design reviews.

The cuts are targeted for the period when the station is being built, with research funding expected to rise again in 2003. NASA officials have promised that past reductions to the research budget would be restored by the time the orbiting laboratory is fully operational in 2004.

Suffredini's memo calls for the number of principal investigators preparing flight experiments to be held "at current levels". NASA had hoped to enlarge the pool of scientists, but the number is now anticipated

to fall by 30 per cent. Contractors are also likely to be laid off "in selected areas".

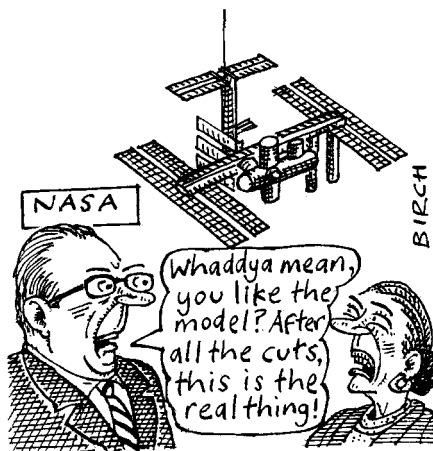
The number of scientists involved has been a "highly visible metric with Congress and the research community", wrote Suffredini. With the proposed cuts, "NASA's commitment to research on [the station] appears questionable".

The idea of raiding the station's research budget, even temporarily, to counter problems caused by the partnership with Russia is sure to raise hackles in Congress. Members of the House Science Committee, which oversees NASA's budget, have repeatedly warned the agency not to short-change science, which has been a leading justification for the project.

Congressional appropriators last year ordered responsibility for the station research budget to be shifted from the human spaceflight office — which operates the space shuttle and is building the station — to the agency's Office of Life and Microgravity Sciences and Applications. But the change has yet to take place.

The reaction on Capitol Hill to a 30 per cent cut, says one congressional source, is likely to be "universally negative". Congress may add more money when it takes up NASA's budget request in the coming year.

Meanwhile, scientists working in the fields of microgravity and life sciences say they are only too accustomed to delays and dwindling budgets for space-station research. When told of the latest funding threat, one NASA-funded scientist quipped: "Ho hum, what else is new?" **Tony Reichhardt**



Biochemistry panel dropped from UK quality review

[LONDON] In a surprise — though widely welcomed — move, Britain's higher education funding councils have removed the biochemistry panel from the list of groups that will carry out its forthcoming Research Assessment Exercise (RAE).

The RAE is an initiative taken every few years to assess the quality of UK research, and is used by the funding councils to provide additional funds for more productive research departments. The next RAE will begin in 2001.

In the past, biochemistry has been assessed separately from the rest of the biological sciences, although this has been seen as illogical by many in both communities. For many years, senior figures had been unsuccessfully pressing for a merger of the panels. Biochemistry departments had been able to take part in the exercise under either panel, giving rise to doubts whether they were being judged on a comparable basis.

Another anomaly was forecast by the Biochemical Society when it began a campaign to have the two panels merged before the last exercise in 1996. It warned that keeping a separate biochemistry panel would lead to only a small number of the best departments choosing to be assessed separately from the rest of the biological sciences.

In line with this warning, in the 1996 exercise, only 17 departments went through the RAE under the biochemistry panel and a notably high proportion of these went on to be rated as 5 or 5*, the top categories in a seven-category rating scale.

According to some researchers, many of these departments appear to have engaged in strategic 'game playing'. By putting a small, high-level biochemistry department through the RAE separately from the rest of a large biological sciences department, it is easier to win top ratings — bringing the hope of receiving extra money.

As recently as last November, after sending a "very focused letter" on the subject, the Biochemical Society thought it had failed to convince the funding councils. The letter urged the councils to remove the panel.

The society has welcomed the change. Details emerged when the Higher Education Funding Council for England released a list of the chairs of unit of assessment panels two days before Christmas. Notably absent was any mention of a biochemistry panel.

"We conducted a fairly extensive consultation last year," says John Rogers, manager of the RAE. "The decision was taken in November to discontinue the biochemistry panel. We are responding to views received to improve the exercise."

Natasha Loder

French scientists offered time to set up companies

[PARIS] French researchers working for public research agencies and universities are to be offered up to six years sabbatical leave to set up their own companies. Scientists would retain their civil servant status during this period and have to choose at the end between a public or private sector career.

The widely expected move was announced last week by Claude Allègre, the minister of national education, research and technology, as part of a package of measures in a bill on "innovation and research" that will go before parliament later this year.

The provisions of the bill also include lifting a ban on scientists, as civil servants, holding shares or sitting on the boards of companies in which they have a direct interest. This has been a major obstacle to researchers aspiring to create their own companies, says Joseph Baexeras, deputy director of industrial development at the Centre National de la Recherche Scientifique (CNRS), France's basic research agency.

A key point of the proposed legislation is that researchers who set up companies would be encouraged to retain close links with their original laboratory, says Baexeras. He points to the absurdity that it is currently illegal for researchers to work for a company that has links with their home laboratory.

Whereas past government industrial policy initiatives have often centred on trying to improve links between public research and the private sector, the proposed law emphasizes the creation of companies as the major means of improving technology transfer and creating high-technology jobs.

Dominique Strauss-Kahn, the industry and finance minister, recently announced a multi-million-franc package of measures to

stimulate innovation. These included FF1 billion (US\$176 million) for new national networks of public- and private-sector laboratories in key technologies, and FF600 million of public funds to boost venture capital (see *Nature* 393, 203; 1998).

The emphasis of the new measures is on identifying and developing promising projects to the stage where they would qualify for venture capital funding. To encourage this, the bill would free industrial development units within the research bodies from the administrative burdens associated with being part of a public body.

Such units would instead be run along the lines of a commercial company, with greater flexibility, for example, in decisions on hiring staff and spending budgets. The creation of joint subsidiaries between research bodies and companies currently requires approval signed by several ministers, but in future such approval would be tacit.

The CNRS intends to create an 'incubator' unit for start-up companies that would provide training in business and finance for scientists. Promising projects would be funded by a series of 'seed money' funds.

The state budget for research this year includes a FF200 million seed money fund which will be distributed following a call for proposals in the spring. Allègre also promised to launch a competition this year with FF100 million to be awarded to the most promising start-ups, in particular in information technology and biology.

The Atomic Energy Commission (CEA) announced last week that it would create a seed money fund of FF160 million for microelectronics. Plans are also afoot to create a fund for biotechnology. **Declan Butler**

US genome researcher strikes it rich

[WASHINGTON] For genome researcher and baseball fan Philip Ozersky of Washington University, St Louis, Mark McGwire's record breaking 70 home runs in the last baseball season was not only an event of semi-religious status, but also highly lucrative.

Ozersky, who attended the game with colleagues from the university, was able to catch the ball, which was sold at auction last week to an anonymous bidder for just over \$3 million — 23 times the previous record price for a baseball.

Ozersky, who attended the auction wearing a tie decorated with a double helix, and says that he will give some of the money to charity, has no plans to give up his work on human chromosomes 7 and 22.



APRON FREHM

India approves use of genetically modified crops, despite critics

[NEW DELHI] Undeterred by criticism from environmental groups, the Indian government announced last week that it will encourage the use of genetically modified seeds in agriculture, and give high priority to transgenic crop research.

It says high priority is necessary because of the "urgency to enhance food production and develop crops with desired traits". But government officials assured critics that the 1998 revised recombinant DNA safety guidelines and biosafety protocols would be fully observed during research and field trials and before modified seeds were marketed.

The announcement came in a statement from heads of agricultural research, the Department of Biotechnology (DBT) and regulatory committees, who held a press conference in New Delhi in the wake of mounting protests against the controversial trials of *Bt* cotton by Monsanto in nine Indian states.

Angry farmers in Karnataka state recently burned experimental fields there (see *Nature* 396, 397; 1998), and the government of Andhra Pradesh stopped the trials. Earlier this month, the Research Foundation for Science, Technology and Ecology (RFSTE), a non-governmental organization in Delhi, issued a lawsuit against DBT in the Delhi high court for authorizing the trial and bypassing the ministry of the environment.

At the press conference, however, the officials claimed the Monsanto trials could lead to an annual saving in the use of pesticides of \$375 million, and accused critics of misleading the environmental groups.

"With 60 million acres of transgenic plants under cultivation worldwide, India cannot lag behind others in this technology," said Manju Sharma, secretary of DBT. She said that DBT is funding transgenic research in tobacco, rice, wheat and potato and has given permission to three other private companies for field trials of mustard, tomato and cauliflower. The \$100 million given by the World Bank to the national agricultural technology project would be used for major research in transgenic crops, say the officials.

According to Asis Datta, chairman of the DBT committee that approved the Monsanto trials, *Bt* cotton was grown in a plot of 25 square metres and then in five plots, each of 200 square metres, before being approved for one-acre fields in 40 locations nationwide.

Vandana Shiva, president of RFSTE, says the trials are illegal because planting started in June last year, before official permission was given in July, and they were not cleared by the Genetic Engineering Approval Committee of the Ministry of the Environment, which has the final say.

K. S. Jayaraman

Physicist's dual role could help with Japanese merger

[TOKYO] Physicist Akito Arima, former president of Tokyo University and currently Japan's education minister, is to take up an additional post as director-general of the Science and Technology Agency (STA).

The appointment of Arima, an outspoken supporter of reforms in the organization of Japanese science, is seen by many as likely to increase the success of the merger of STA and the Ministry of Education, Science, Sports and Culture (Monbusho) in two years' time (see *Nature* 390, 327; 1997).

It follows the departure last week of three cabinet ministers, including Yutaka Takeyama, the current director-general of STA, as part of a cabinet reshuffle at the launch of a coalition between the ruling Liberal Democratic Party (LDP) and the Liberal Party.

The leaders of both parties reached an agreement on 13 January, ending a series of talks aimed at ironing out policy differences, especially on security issues. The coalition government was launched on 14 January in time for the start of the ordinary session of the Diet, Japan's parliament, on Tuesday (19 January).

Keizo Obuchi, Japan's prime minister and the LDP's leader, reduced the number of ministers from 20 to 18 by merging several ministerial posts, including those for STA and Monbusho, as well as those for the Ministry of Construction and the Land Development Agency.

Despite speculation that Arima, who is relatively inexperienced in politics, would be the first to leave the cabinet, Obuchi instead asked for voluntary resignations from ministers of his own political faction.

After resigning as director of STA's Institute of Physical and Chemical Research (RIKEN) last year, Arima was elected to the upper house of the Diet on LDP's proportional representation list (see *Nature* 394, 512; 1998) and, just two weeks later, was appointed education minister in Obuchi's new cabinet.

He introduced the first external review system to a Japanese university and helped draft the 1996 Basic Law for Science and Technology, designed to increase Japan's spending on science by 50 per cent by 2001.

Japanese scientists have shown strong support for Arima's appointment as the head of STA, especially as many were concerned about its incompatibility with Monbusho, which has an education-orientated, 'bottom-up' approach to research, as opposed to STA's 'top-down' approach. They hope that Arima's experience of both worlds will enable him to help resolve conflicts between the cultures of the two agencies.



Arima: 'at home with STA and Monbusho'.

Institutes under Monbusho are worried that STA will drive their research towards application-orientated projects. STA, which is smaller and less powerful than Monbusho, is afraid it would be affected by Monbusho's relatively inflexible funding system.

"I cannot think of anyone but Arima who could fulfil the role as head of STA and Monbusho" says Minoru Oda, former director-general of Monbusho's Institute of Space and Astronomical Science, who has also headed RIKEN. "Scientists are often ignorant about politics, and politicians often fail to understand the importance of scientific research — he is the only person with an insight and understanding of both ends of the spectrum."

Oda expressed concern that the merger of STA and Monbusho could jeopardize support for basic research, especially if the new ministry has a bureaucratic approach.

At a press conference last week, Arima said he felt "at home with both STA and Monbusho" and hoped "that the dual post would help promote science and technology in Japan".

STA and Monbusho have started to form detailed plans for the impending merger, which will involve a drastic reduction in the number of sections on both sides. According to its preliminary plan, the new ministry will have seven bureaux, including those for science and technology policy, promotion of research, research and development, and higher education.

STA's role in nuclear safety management will be reduced by the merger: the main part of its nuclear safety bureau, responsible for regulating nuclear materials, will be transferred to the Ministry of International Trade and Industry, and a section responsible for regulating reactors and radioactive material for basic research will be transferred to the new bureau of science and technology policy.

The atomic-energy bureau will be reduced to sections within the two bureaux of the promotion of research, and science and technology policy, respectively.

Large projects, such as Monju, the experimental fast-breeder reactor, will be overseen by the former, with smaller projects, such as those concerning radiation research, moving to the latter.

Asako Saegusa

Fossil dealer charged over Russian cache

[MUNICH] Russian prosecutors are hoping that the arrest of a German fossil dealer prevented from taking fossils and meteorites out of Russia may help them to track down a group they suspect of being involved in the theft of fossils from Russian museums.

Joachim Wördemann was stopped on 21 December trying to drive an estimated 1,000 kg of large meteorites, minerals and fossil material over the border into Finland. He appeared in a St Petersburg court on 5 January charged with irregularities in export documentation and failure to provide export licences for around ten per cent of the items.

Although most of the items had export licences issued by the Ministry of Culture, these licences are being investigated by police. Many were countersigned by directors of the

Institute for Palaeontology in Moscow, part of the Russian Academy of Sciences, which officially advises the ministry on the scientific value of fossils intended for export. Fossils identified as scientifically important are not eligible for export.

The directors of the institute have been strongly criticized by Russian palaeontologists for operating an opaque system of administration that scientists in the institute claim often conflicts with its scientific interests (see *Nature* 384, 499; 1996).

Moreover, the directors have been caught up in a scandal over the theft of many important fossil specimens from the institute's museum, which is now the subject of a criminal investigation (*Nature* 391, 724; 1998).

In 1994, Wördemann himself was stopped

from selling a fossil originating from the institute — a *Thoosuchus jacovlevi* skull worth US\$950 — but he denied being involved in the theft and police lacked sufficient evidence to bring a prosecution (see *Nature* 371, 729; 1994). Following his recent arrest, German police are once more investigating his past activities.

At the request of the Russian ministry of the interior, the police are gathering information from collectors and museums about any contacts they may have had with Wördemann. Wördemann's Russian fossils have, until last year, been for sale — cash only — at fairs and privately in Germany since 1989, says Rupert Wild, head of the palaeontological department of the State Museum for Natural History in Stuttgart.

Wördemann visits Russia twice a year and is a regular visitor to the palaeontology institute. After his arrest in St Petersburg, he was sent to Moscow where he spent several days answering questions from police; this was believed to be in connection with the investigation into the situation at the institute.

Meanwhile, customs officials called in a scientist from the Zoological Institute in St Petersburg to check whether Wördemann's collection included a mammoth tusk that had recently gone missing from the institute; it did not. Other institute scientists expect to be asked for advice on the scientific value of fossils in the load during the next few weeks.

Items likely to be considered scientifically important include complete baby mammoth skulls from the Southern Urals which show developing teeth, and many cave bear skeletons and skulls. The load also includes mammoth tusks, ammonites and trilobites of unknown scientific value. The total market value in the West could be millions of dollars.

Scientists at the palaeontology institute have long suspected that the fossils disappearing regularly from their museum were being smuggled out of the country by a criminal network whose members, they deduced, appeared to include both institute members and customs officials. Four years ago they pinpointed the customs office through which they were probably being funnelled; this was the same office at which Wördemann was arrested.

Afraid to talk in case they lost their jobs, and failing to win support from the Academy of Sciences, the scientists eventually managed to interest politicians, who initiated an investigation by public prosecutors.

In a television broadcast, the national public prosecutor said the Russian authorities were cracking down on organized crime rings, which, he said, were stripping Russia of her national heritage. He said Wördemann's arrest was the first time police had stopped an apparent smuggler at a border. **Alison Abbott**

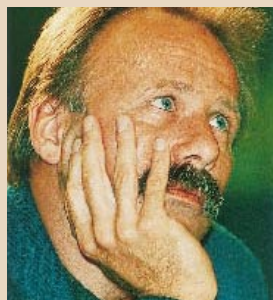
Greens accept reprieve for neutron source

[MUNICH] The German Social Democrats (SPD) last week blocked a move by the federal environment minister Jürgen Trittin, a leading member of the Greens, to stop the licensing of new research reactors with a thermal power of more than 1 MW.

Trittin had specifically called for the termination of the licensing procedure for FRM II, the 20-MW neutron source being built by the Technical University of Munich at Garching (see *Nature* 397, 92; 1999).

But the Greens and Social Democrats agreed during last week's coalition talks on nuclear policy to exclude research reactors from proposed legislation on the phasing out of nuclear power. They also agreed to set up an independent committee of scientific experts to reconsider whether the FRM II could be converted to use low enriched uranium (LEU).

The FRM II has been designed to burn weapons-grade highly enriched uranium (HEU), contravening a 1978 international anti-nuclear proliferation agreement. The decision not to block the licensing of such reactors was described as "a bitter pill" by Harald Händel, spokesman of the Green



Trittin: his attempts to block FRM II seem to have failed.

party's executive. Party leaders, he says, are aware that the grass roots will have little sympathy for the compromise: "but at the end of the day it was all that was possible politically."

Wolf-Michael Catenhusen (SPD), secretary of state in the federal research ministry, welcomes the compromise, which, he says, "guarantees that Germany will be able to keep up its options in neutron research". But he stresses that the political desire to convert the FRM II to use LEU remains.

The new committee will have until June to formulate recommendations about whether conversion is "technically and financially feasible". The federal government will then discuss the reactor's final design with the Bavarian state government, which is

responsible for issuing the operating licence.

The composition of the committee has not yet been decided. Catenhusen told *Nature*, however, that it will consist of prominent neutron researchers, including experts from the Technical University of Munich.

He says it will also include experts from the US Argonne National Laboratory, who last year published a report saying that they saw "no major technical issues regarding use of LEU fuel instead of HEU fuel in the FRM II", and "that it is definitely feasible to use LEU fuel in the FRM II without compromising the safety or performance of the facility".

Meanwhile, the Technical University of Munich is keen to present its critics with a *fait accompli* by ensuring that FRM II building work continues as fast as possible, daily increasing the costs of redesigning the reactor.

Armando Travelli, head of Argonne's reactor conversion programme, fears such costs could be "immense", and that this could be used as an argument not to convert the reactor. But Catenhusen has indicated that the federal government would contribute to the conversion costs. **Quirin Schiermeier**

Plan to limit Cambridge high-tech growth

[LONDON] Local authorities in the east of England are preparing to defend controversial plans to restrict the numbers of new high-technology companies setting up near the city of Cambridge.

The plans are being opposed by business organizations which want new information technology and biotechnology companies to be located on the outskirts of Cambridge near existing clusters of such businesses, as well as the university and the science parks.

The newly formed East of England Business Group is to challenge the proposed planning restrictions at a public examination starting on 2 February. Last month, a report commissioned by the group of 10 business organizations, which includes the

Confederation of British Industry and the Institute of Directors, said restrictions would damage prosperity and economic growth.

Written by the Cambridge-based management consultancy Segal Quince Wicksteed, the report says that planning restrictions will conflict with a recent government white paper, which promised to find ways of supporting clusters of high-technology industries (see *Nature* 396, 714–715; 1998).

But the Standing Conference of East Anglian Local Authorities (SCEALA), which covers the three counties of Cambridgeshire, Norfolk and Suffolk, says it wants to restrict the pace of development in Cambridge's existing technology clusters, and to regenerate other, more deprived urban areas.

Authorities belonging to SCEALA are concerned that the region has too few skilled workers to meet continued growth in Cambridge's high-technology sector. Staff would have to be recruited from outside the region, leading to greater pressures on roads and housing, which the counties wish to avoid.

The Cambridge region is home to more than 1,200 high-technology businesses, compared to 350 in the mid-1980s. It is one of Britain's most prosperous regions. Between 1991 and 1996 it accounted for all regional growth in employment. Some parts of Norfolk and Suffolk, by contrast, have employment rates and incomes per head below the UK average.

The previous Conservative government

Quebec's tax incentives act as a magnet to research companies

[MONTREAL] Tax incentives and other measures introduced in Canada's budget last March appear to be turning Quebec into the country's most attractive province for high-technology companies that carry out research and development (R&D).

Most other provinces also have generous R&D tax subsidies for manufacturers, making the average rate for Canada the highest in the industrialized world. Defining tax subsidy rates as the net cost of investment after taxes and subsidies puts Canada's rate at more than twice those of Australia and the United States, the next highest.

But for foreign-controlled non-manufacturing firms in particular, Quebec's corporate tax rate of 9.15 per cent is well below the average for the other nine provinces of 16.1 per cent. When added to the savings afforded by Canada's weak dollar, this makes Quebec the cheapest place in North America to conduct research.

During the past seven months, new institutes have opened in Montreal, Quebec City, Hull and Laval, where new (and not-so-new) companies can locate and gain benefits including a five-year holiday from income, capital and payroll taxes.

In their first three years there, companies can obtain tax credits of up to 40 cents for each Canadian dollar spent on wages for eligible employees, and on the purchase or lease of equipment. The institutes have advanced telecommunications equipment and infrastructure.

Vorton Technologies, for example, a software company based in Ontario, found that by moving about 25 kilometres across the Ottawa River to Hull, in Quebec, it could save more than Can\$100,000 (US\$65,000) a year through tax benefits.

The saving is enough to pay the rent in one of the new institutes. The company will



Irresistible offer: tax breaks are luring companies to Montreal (above) and other Quebec cities.

also be able to hire more professional staff, so it can do more R&D, according to Vorton's chief technology officer, Tony Davidson. Many more companies are considering making similar moves.

Over the past three years, seven other provinces have offered tax credits, and British Columbia plans to do so. Ontario's R&D tax incentives total Can\$100 million.

But the tax subsidies have their critics. In a report last September, the Organization for Economic Cooperation and Development (OECD) said such subsidies should be "rationalized". It said tax breaks should be incremental, rising with the level of R&D performed rather than offering a set value per dollar of R&D investment.

The report says that: "A comparison (between) Australia, Canada and the United States concludes that volume-based tax incentives do not generate much R&D beyond the tax expenditure incurred by the government".

One study it cited of 55 Canadian firms found that the value of R&D performed "did not amount to more than 40 per cent of lost tax revenues". Incremental tax credits, on the other hand, have generated as much as \$2 in R&D for every dollar in tax credits.

Canadian officials contest this, however, claiming that their R&D tax incentives have sometimes generated more than three times the value calculated by the OECD. Pierre Etienne Gregoire of Quebec's Ministry of Economy, Industry, Science and Technology, says Canada has tried incremental tax credits, but results were not good.

Jack Mintz, professor of business at the University of Toronto and author of a major study of corporate taxation, says the Canadian tax subsidies are not working. Canada's ratio of private R&D expenditure to gross domestic product (about 1 per cent) is much lower than that in Sweden (2.3 per cent), for example, as well as those of other countries with smaller populations.

"The question is whether the tax system is the best way to deliver all the R&D assistance that we expect," says Mintz. He also thinks that Quebec's use of the new low-tax institutes will not be the answer.

The institutes may result in more R&D being performed there, he argues, but they will not necessarily provide a proportional commercial spin-off or job creation.

The political uncertainty of locating in Quebec, whose Parti Quebecois government is committed to separating from Canada, does not seem to deter companies from moving there.

David Spurgeon



Prime location: the Napp Building, one of the research facilities at Cambridge Science Park.

tried to ease the pressure on Cambridge, and to reduce the regional economic disparity, by encouraging high-technology businesses to move to poorer rural areas. These would be linked by new roads to major inter-city roads, enabling staff to commute.

The new Labour government has taken a different approach. It cut the roads programme last year. And the SCEALA authorities, which have changed from being predominantly under Conservative control, to Labour, have decided to concentrate on urban — rather than rural — regeneration. In addition, a government committee chaired by architect Lord Richard Rogers said last week that new housing should be built on disused 'brownfield' manufacturing sites in urban areas and not on 'greenfield' rural sites.

Under SCEALA's proposed planning restrictions, new businesses would be encouraged to locate in or near towns and cities other than Cambridge. The strategy also aims to encourage greater use of public transport for travel to work.

But Bill Wicksteed of Segal Quince Wicksteed says that preventing the growth of clusters of high-technology businesses will ultimately disadvantage local communities as well as the high-technology industry itself.

"Growth should not be grudgingly accepted, but embraced and shaped," says Wicksteed. "If [local authorities] work from an assumption that growth must be resisted, they will fail. They will get growth, but they will be unable to control it."

The Wicksteed report believes that it is inappropriate for SCEALA to try to integrate its technology management policies with the aim of urban regeneration. And it doubts whether technology companies will want to move to areas where they may not be able to find enough qualified staff.

The controversy will intensify the government's difficulties over a planning application from the Wellcome Trust for a genomics research complex on the outskirts of Cambridge, which the local authority opposes. Prospects for the complex will not have been helped by the recent resignation of industry secretary Peter Mandelson, the architect of the government's decision to promote knowledge-based industry. **Ehsan Masood**

Tough times set to continue for US biotech start-ups

[SAN FRANCISCO] US biotechnology companies are bracing themselves for continued financial problems in 1999 as the public markets look to industries that can provide quicker, more reliable returns on their money, according to industry analysts.

Those attending the annual Hambrecht & Quist (H&Q) healthcare investment conference in San Francisco last week were also told that, although investing in high-risk venture companies is likely to reach an all-time high in 1999, biotechnology is unlikely to enjoy the benefits.

Industry observers did not believe that the fall in funding would trigger a slowdown in science. "There's such innovation in biomedical research that, for every failure, there are probably ten professors writing a business plan," says Alex Zisson, pharmaceutical analyst for the investment company.

But even top university administrators acknowledge that outside interest in licensing products or funding start-up companies from bioscience research has slipped. Katherine Ku, director of the office of technology licensing at Stanford University, says that entrepreneurial activity in the physical sciences has outstripped biotechnology for several years.

About 20 biotechnology companies, including some long-established ones, slashed operations or closed their doors entirely last year due to the cash crisis. Casualties included Alpha One, Biocircuits, Cellex, Cellpro, ChemTrack and ImmuLogic, and more failures are predicted this year.

"Too much capital is needed to bail everybody out," says Dennis Purcell, managing director and head of life sciences banking for H&Q. He predicts that smaller companies will find it hard to raise the money they need.

According to Purcell, about 95 biotechnology companies have less than one year of cash available to fund operations. The industry raised only \$951 million in public shares last year — a 60 per cent drop from the \$2.28 billion collected in 1997. Initial money raised in shares was \$417 million in 12 companies launched on the stock market, 44 per cent down from the \$750 million collected the year before in 22 flotations.

Kurt Von Emster, portfolio manager for the Franklin Biotechnology Fund, attributes much of the crisis to the high level of enthusiasm in 1991 and 1992, which fuelled a boom of marginal companies and products. But poor market performance for six years running and the largest ever number of projects in which drugs failed to fulfil their promise have created a desperate need for capital, he adds.

Von Emster estimates that about a quar-

ter of public biotech companies need to raise money, while another 100 are still waiting to go public after six years of private operations. "About \$5 to \$8 billion need to be raised over the next 16 to 18 months," he said. "That's just not going to happen."

Even though the Nasdaq biotechnology shares index climbed 40 per cent over the year, representing a sharp recovery from falls of 2 per cent in 1997 and 1 per cent in 1996, most of the gain came from the top ten companies. Almost two-thirds of publicly quoted companies in the sector lost value, and 50 are trading below their cash value, says Purcell.

But he is optimistic that investors will return, lured by the gap between low stock valuations and the value represented by product potential and real revenues.

Venture capitalists at the H&Q conference said that, while funding is more difficult to come by these days, good ideas will not go begging. They seemed most charmed by promises of steady revenue from young companies that could sell subscriptions or service contracts for platform technologies such as genomics or drug-target validation.

Zisson points out that, although enthusiasm for technology transfer in the academic world remains high, in today's climate a great technological product is not enough: scientists must come up with a smart business plan as well.

Lita Nelson, director of technology licensing at the Massachusetts Institute of Technology, says that, as leading venture companies have become consolidated, this has left a gap at the level of creating companies. The pioneer funds have become so successful that they cannot afford to make smaller investments or put the time into helping small companies get started.

While a few new venture companies have moved in, start-ups have lost the benefit of the older, established venture capitalists' experience, say Nelson and others. These young companies will probably have a hard time surviving entrepreneurial speed bumps, such as raising more capital when needed. "These companies won't be as successful when they run into trouble," says Nelson.

Jeff Casdin, chief executive of Casdin Capital Partners, is optimistic that there will eventually be a broad recovery in the value of biotechnology shares. He predicts that the scientific potential will become irresistible to investors.

"If the stuff is good, interesting, kicking, I don't think it's hard to get attention," says Jan zur Hausen, an associate with MPM Asset Management of Cambridge, Massachusetts. **Sally Lehrman**

Industry and research meet health service to set up hormone centre

[LONDON] A “unique alliance” between academia, industry and Britain’s National Health Service (NHS) was forged last week with the launch of the Oxford Centre for Diabetes, Endocrinology and Metabolism. Its “care to cure” approach is intended to allow patients faster access to medical research.

The £10 million (US\$16.5 million) centre will be built with funding from the pharmaceutical company Novo Nordisk and the NHS, and will provide a base for Oxford University’s medical research and clinical treatment in these areas. The partnership between patients and research is intended to accelerate innovation in treatment.

Launched by the Secretary of State for Health Frank Dobson, the centre is claimed to be the first in the world to integrate basic and clinical research, clinical care, and scientific and patient education in hormone-related and metabolic diseases.

German science budget rises, but not by much

[MUNICH] Germany’s budget for science and education is expected to rise by DM1 billion

(\$590 million) to DM16 billion when the 1999 federal budget is approved this week. This is not enough to allow Social Democrat science minister Edelgard Bulmahn to keep her election promise to double federal spending on research within five years.

But the increase is still seen as a victory at a time when other ministries are being asked to tighten their belts. Most of the additional money will be spent on Germany’s underfinanced university building programme. But Bulmahn is also likely to announce new research programmes in molecular medicine and information technology.

‘Agriethics’ committee set up in France

[PARIS] After bioethics comes ‘agriethics’. Spurred by controversies over ‘mad cow’ disease, genetically modified crops and cloning, Guy Paillotin, director general of the French national agricultural research agency, INRA, has decided to set up an ethics committee to produce ‘opinions’ and recommendations on issues raised by agricultural technology.

The committee will fill a gap left by existing bioethics bodies by reflecting on the complex relations between public opinion, food and farming practices, says Paillotin,

whose book *Shut Up and Eat* is due out next month. The committee will be chaired by Jean-François Théry, a legal expert who was director general of the science ministry in the early 1980s, and its members include neurobiologist Jean-Didier Vincent and biologist Hervé le Guyader.

Public understanding of science wins £3m boost

[LONDON] The British Association for the Advancement of Science is to receive a grant of just under £3 million (US\$4.9 million) from the Wellcome Trust to support its efforts in promoting the public understanding of science.

The grant will be allocated over five years and used to help the association implement its future strategy. This is to include providing focal points for science communication across Britain. “Now, perhaps more than any other time in our history, science needs debate,” said Michael Dexter, director of the Wellcome Trust.

Bovine growth hormone sales banned in Canada

[MONTREAL] Canada’s federal health department has said that it will not approve the sale of the bovine growth hormone

recombinant bovine somatotropin (rbST) after a panel of veterinary experts found it increased health risks to dairy cows. However, another panel found that the use of the hormone posed no significant risk to humans.

Officials said the decision was based on more than nine years of comprehensive review. The drug is intended to increase milk production in lactating cows by 10 to 15 per cent. It was approved for use in the United States in 1994 and declared safe by the European Union in 1990, although a moratorium was placed on its sale in EU countries until the end of the decade.

Astronomers win fight to keep Arizona in the dark

[WASHINGTON] County supervisors in Tucson, Arizona, have blocked a planned commercial development that astronomers say would have ruined dark-sky viewing conditions on nearby Mount Hopkins. The four-to-one vote by the Pima County Board of Supervisors to refuse permission for commercial development at Canoa Ranch was a rare defeat for developers in southern Arizona (see *Nature* 397, 95; 1999).

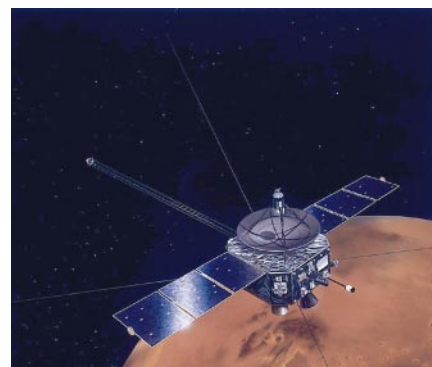
It was also an important victory for local astronomers, who had been threatened with

legal action by Fairfield Homes, the Canoa Ranch developer, if they spoke out against the project. More than 2,200 astronomers from 30 countries had signed a petition opposing Fairfield's tactics and the threat to viewing conditions on Mount Hopkins.

Editor axed for 'political' survey of oral sex

[WASHINGTON] The American Medical Association last week ousted the editor of its journal, *The Journal of the American Medical Association (JAMA)*, claiming that he tried to influence the course of the Senate trial of President Bill Clinton. The editor had published an article in which only 41 per cent of college students surveyed say that engaging in oral sex amounts to "having sex".

The association's executive vice-president, E. Ratcliffe Anderson, accused George Lundberg, the outspoken editor of *JAMA* for 17 years, of "inappropriately and inexcusably interjecting *JAMA* into a major political debate that has nothing to do with science or medicine". Lundberg's lawyer said in a statement that the association had chosen to "jeopardize the editorial integrity and scientific credibility of [*JAMA*] for political ends".



Japanese probe misses first chance at Mars

[TOKYO] Japan's first Mars probe, Nozomi (above), will enter the Martian atmosphere four years later than originally planned. It is thought that a failed 'swing-by', intended to take the craft out of its elliptical orbit around Earth, has exhausted the fuel supply needed for the 700 million kilometre journey.

The Institute of Space and Astronautical Science said last week that Nozomi will orbit the Sun three times before attempting another swing-by to launch its journey to Mars. Nozomi is now scheduled to arrive at Mars around the end of 2003 or the beginning of 2004, where it will spend two years analysing the planet's atmosphere.



The writing is on the web for science journals in print

The Internet revolution is injecting more competition into publishing and giving power back to scientists and learned societies. It presents new challenges to the guardians of the archives and could yet spell the end for many print titles.

Print versions of scientific journals will soon be history at Denmark's national Technical Knowledge Center and Library in Lyngby. It has decided to phase out print altogether, and deliver journals direct to staff desktops via the World-Wide Web.

In itself this move, expected to be increasingly followed by universities worldwide, is a revolution. But it is only one of the profound changes in scholarly publishing that are just around the corner. The very existence of research libraries as we know them is in doubt.

Their traditional roles are being eroded on every front. Publishers and new electronic services are bypassing libraries and delivering sophisticated information products direct to users. Libraries increasingly act merely as brokers to squeeze better deals from publishers.

Publishers face challenging times ahead too. The late Robert Maxwell built his empire on the lucrative business of publishing journals, many of them obscure titles read by a few. But the journals system is increasingly looking like a house of cards, and electronics is shaking the foundations.

Much evidence suggests that maintaining a plethora of high-price, low-circulation journals as the primary means of scientific communication is no longer the best way to

meet the needs of users, and that the overheads are economically unsustainable.

The ability to click from an abstract or citation to the full text of an article is prompting a shift in the way that journals are used. Scientists often care less about the journal title than the ability to track down quickly the full text of articles relevant to their interests. Increasingly, users view titles as merely part of hyperlinked 'content databases' made up of constellations of journal titles.

As a result, the competitive edge of publishers is increasingly coming to depend on their ability to muster a critical mass of attractive information through a single powerful and user-friendly interface. The boundaries created by thousands of journals appear as little more than an evolutionary vestige. The big bang of journal proliferation seems set to be followed by a big crunch.

There is growing acknowledgement that the primary role of journals will in future be to provide papers with an imprimatur of quality and to add editorial value. Their traditional role as a distribution outlet will diminish.

This is already happening in physics, where the Los Alamos e-print archives have become the primary means of communication, not just in high-energy physics but also in astrophysics, quantum physics, condensed matter theory, mathematics and computer science. Plans to create similar archives for

the biomedical sciences are receiving serious attention in the United States.

The Internet is blurring the traditional roles of creators, suppliers and distributors of scientific information, and injecting a long overdue element of competition. A shake-out of the entire scholarly publishing industry seems inevitable. "It is an interesting time. The publishers don't know which way to go, and the libraries don't either," says David Lipman, director of the US National Center for Biotechnology Information.

Serial library killer

Going fully electronic was the only way for the Danish library to continue to provide an efficient and cost-effective service, says Lars Bjoernshauge, its director. Like research libraries worldwide, it has been badly hit by the spiralling inflation of journal prices that has resulted in libraries paying more to provide users with less — the 'serials crisis'.

The US Association of Research Libraries calculates that its 114 member libraries spent 142 per cent more on journals in 1997 than ten years before, but ordered 6 per cent fewer titles. In the same year, Reed-Elsevier, one of the largest publishers, reported profits of £230 million (US\$378 million) on sales of £571 million in its scientific activities alone.

"Between 1990 and 1997, we were forced to cut 40 per cent of our subscriptions," says Bjoernshauge. The Danish library has shifted resources away from handling paper — unpacking, cataloguing and shelving journals — and fired one in seven staff. As a result, this year it expects to increase its titles by 25 per cent. "We now have more journals, less staff, and satisfied users," he says.

Just four years ago, such a move would have been unimaginable. The total number of journals, of any sort, on the web was just 306. Even 18 months ago, it would have been unthinkable, as only a few scientific journals had full text and graphics on the Internet.

The major change since then has been the arrival of traditional publishers on the web. Reed-Elsevier now has more than 1,200 journals online, Springer has 360, and Academic Press 174. A journal without a web version is now rare, and probably endangered.

The Danish action comes as no surprise to Andrew Odlyzko, a mathematician at the AT&T telecoms corporation and an expert on the economics of scholarly publishing. He has long argued that the 'journals' crisis is part of a wider crisis of identity facing research libraries, pointing out that the cost of journals is typically only one-third of a library's total staff and overhead costs.

Few other libraries are yet prepared to abandon print completely. The main obstacle to faster change is concern over how electronic materials are to be archived for posterity. Universities that have contemplated

This Briefing has been written by Declan Butler

On other pages | Libraries fight back: p196 | Archiving challenge: p198 | No more inky fingers: p200

cancelling print subscriptions have frequently met resistance from faculty members. But a marked shift of resources from print to electronics is already under way as libraries take advantage of electronic publishing to cut overheads. And many scientists hope that electronics may bring a more fundamental reform of the economics of scholarly publishing, and a solution to the 'serials crisis'.

Many feel that publishers of high-price, small-circulation journals are making excess profits. One critic is Mark McCabe, an economist at the Georgia Institute of Technology in Atlanta, who spent seven years at the US Department of Justice's antitrust division investigating anti-competitive practices. He says the profit margins enjoyed by many publishers exceed those that would be expected in a properly competitive market.

Commercial publishers argue that they have higher costs than not-for-profit societies, whose journals are often subsidized by membership fees, and that increases in the quality and size of journals have raised prices. But McCabe maintains, for example, that it is difficult to justify the doubling of the cost of Elsevier's *Brain Research* between 1992 and 1996 to \$15,000 annually. Elsevier Science was unavailable for comment.

McCabe: keen to see more competition.

Libraries strike back

Libraries strike back

Libraries and scientists are now striking back, the ultimate goal being to return control of scholarly publishing to the non-profit societies and what they consider to be responsible publishers. Using electronic journals to achieve this is the goal of HighWire Press, a not-for-profit outfit set up in 1995 by Stanford University Libraries and Academic Information Resources to help universities and societies to publish at low cost.

Michael Keller, publisher of HighWire, says he was concerned that large companies with more money to invest would squeeze not-for-profit publishers out of the electronic journals market. He hopes that HighWire will help "correct the market", by increasing the output and quality of society journals. "This market problem has taken 40 years to come about; my guess is it will take ten to fifteen years to remedy."

HighWire now has more than 100 journals in its stable, including the *Journal of Biological Chemistry* — the world's most cited journal — *Science* and *Proceedings of the National Academy of Sciences*. Just before Christmas, Oxford University Press transferred responsibility for production and hosting of electronic versions of 160 journals to HighWire.

The Scholarly Publishing and Academic



Gateway to literature: ISI's Web of Science takes readers to the full text of more than 100 journals.

Resources Coalition (SPARC), set up in 1997 by the US Association of Research Libraries (ARL), is even more aggressive. It is underwriting the launch of journals aimed at competing head on with expensive titles, with its 114 member libraries promising to buy each of them (see *Nature* 393, 719; 1998).

SPARC has teamed up with the UK Royal Society of Chemistry to launch an electronic journal, *PhysChemComm*, that will sell at \$353 and is intended to compete with Elsevier's \$8,000 *Chemical Physics Letters*. But Elsevier complains that like is not being compared with like, and that another of its journals, *Electrochemistry Communications*, sells at \$350.

The most dramatic example of rebellion is perhaps the recent decision by Michael Rosenzweig, a researcher at the University of Arizona, to defect, along with the entire editorial board, from the Wolters Kluwer journal, *Evolutionary Ecology Research*. Rosenzweig had become disenchanted with price

increases at the journal which he established 12 years ago, and has allied with SPARC to create an alternative that will sell to institutions at around one-third of the \$777 price of the Kluwer journal.

"I think the broad support that SPARC has received speaks of a broad frustration among [researchers] with the situation and a belief that it is time to become part of the solution," says Mary Case, director of ARL.

Market distortions

The success of such electronic ventures is far from guaranteed, however. Despite enjoying library support, electronic journals face similar difficulties to those of any new journal in getting established and attracting authors.

Gregory Fu, a chemist at the Massachusetts Institute of Technology, also points out that, from a user's perspective, "unless *PhysChemComm* delivers a knock-out blow [to its competitors] it just becomes one more journal that I have to put on my list to browse; so long as the Elsevier journal has good papers I will still need to check it".

Case admits that her optimism is tempered: "If new SPARC titles can draw papers away from top expensive competitors, there should be a decrease in size or quality and eventually a decrease in price. There may be some modest effect on prices, as much due to the publicity that SPARC is generating as to the introduction of competition."

In physics, the Los Alamos e-print repository — where preprints submitted direct by

User consortia emerge as 'brokers'

In addition to efforts by scientists to create competing products, publishers face the threat of more organized action — and perhaps even a boycott — by consortia of libraries and other users intent on forcing down prices charged for electronic content.

Library consortia, established in the 1930s to cooperate in administering interlibrary loans, have over the past two years taken on a new role: squeezing better deals out of publishers for electronic licences. The movement is becoming international; the International Coalition of Library Consortia (ICOLC), set up in 1997, groups 79 library consortia in North America, as well as a growing number in other countries, including the United Kingdom, Germany and Australia.

Two dramatic examples are Ohiolink, a consortium of 74 Ohio libraries, which negotiates state-wide access and provides a common interface to users, and the new National Electronic Site Licence Initiative (NESLI), which is intended to serve Britain's entire higher education system.

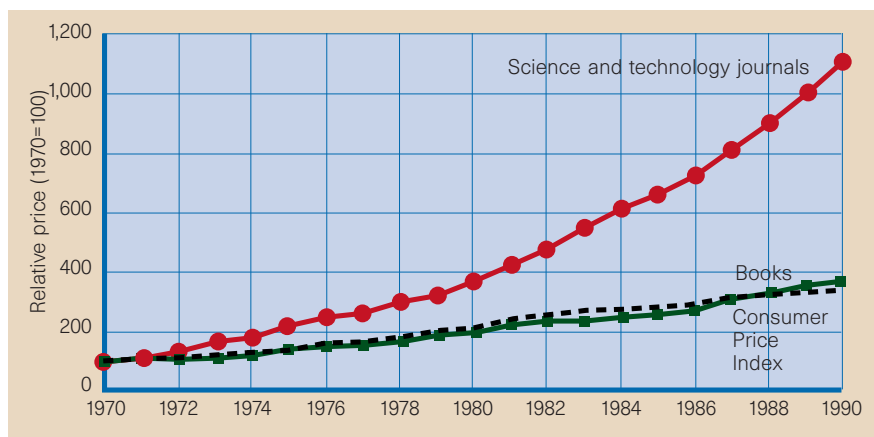
NESLI was created by the Joint Information Systems Committee, a body established by Britain's higher education

funding councils to coordinate the use of information technology by the higher education and research communities. The contract to manage NESLI has been won by a consortium led by the University of Manchester and Swets and Zeitlinger, one of the world's largest intermediaries in the distribution of scientific information. Their brief is "to lessen the financial, legal and technical barriers to the take-up of electronic journal provision... and negotiate value for money deals with publishers".

Paul Harwood, director of Swets and Zeitlinger UK, says NESLI has approached 60 publishers and will begin making electronic journals available this month. "Any [financial] impact will only be seen in subscription years after 1999," he adds.

"NESLI is an attempt to offer a complete service for the UK higher education community, from negotiation to access," says Harwood. He says the initiative will "try to bring some order to a chaotic world by offering a single service for the complex and time consuming aspects of arranging licences and access to electronic journals".

California State University has taken this approach a step further, and turned the



Serials crisis: journal price inflation is bad news for libraries but is giving a push to web publishing.

scientists are automatically made available free — now serve tens of thousands of users worldwide and process millions of electronic transactions per month. In many fields of physics, they have supplanted journals in distributing primary literature. Yet the repository has not dented the financial health of physics journals.

Although this is partly because journals satisfy a demand for peer-reviewed collections, it also reflects what many see as distortions in the publishing market. Scientists depend on publishing for career advancement, but they do not pay directly for journals, so have no incentive to stop submitting to high-priced titles. And, as long as publishers attract good authors, libraries will come

under pressure to buy journals, some of which they cannot afford.

ARL data show that library spending per US faculty member averages \$12,000 per year. "If researchers internalized the price of journals in their budgets they would behave differently," argues McCabe.

The lack of direct accountability of researchers for publication costs is also inhibiting one promising model of Internet publishing, based on billing authors page charges to publish, and then making the journal free. Authors have little incentive to pay page charges when they can publish elsewhere for free.

The viability of the model has been demonstrated by one electronic journal, the

tables on vendors of scientific journals and databases. Just before Christmas, it announced that, instead of negotiating individual deals, it would put out to competitive tender a contract for the building of a database to supply all its 22 campuses with 1,300 specified journals. Like Ohiolink, it also aims to provide a common database and interface to electronic journals for all its students and faculty members.

Its action is partly a response to the practice of many publishers and intermediaries of selling licences to all their electronic journals, or large bundles of them. Many libraries are concerned that as a result they will be required to pay for journals they do not want. "As the bundled database grows, there will need to be concomitant price increases," argues Mary Case, director of the US Association of Research Libraries. "I am not convinced that such increases will be moderate."

Case believes that it would be impossible under US law for university-based consortia to go so far as to organize a boycott of expensive journals. "Such organized activity is illegal," she points out, as it would constitute antitrust practices.

But Mark McCabe, a former official at the US Department of Justice, says the department might grant immunity to pursue such action if libraries were to persuade it that US antitrust guidelines give companies excessive control of the publishing market.

"On occasion the department develops industry-specific guidelines that acknowledge the peculiarities of a market," says McCabe. "Procedures exist for exceptions in areas which for public policy reasons deserve immunity". The department's interest in the scholarly publishing market has been aroused, he adds, and there is increasing recognition that current antitrust regulations may be poorly suited to tackling the quasi-monopoly enjoyed by some publishers.

Even without such drastic action, Case believes that much can be done by appealing to researchers: "The real success will come if authors and editors look at the titles they support, and refuse to sit on editorial boards, submit to and review papers for some of these expensive titles. This is the dramatic change that has to take place to iron out the distortions in this market."

Florida Entomologist, which charges authors \$45 per printed page and \$20 for each figure or table. Thomas Walker, its editor, says adoption of page charges by other professional societies would allow inexpensive journal production, while making information more widely available.

At a broader level, a grassroots movement has emerged over the past year — mainly in the United States — whose goal is to challenge the common practice whereby publishers retain copyright of articles and forbid reuse of the work elsewhere. A compromise that gives researchers greater control of their published work now seems inevitable.

Some publishers have begun relaxing the terms of copyright agreements to allow researchers to resubmit articles to other media, such as digital libraries. Several US universities, including Caltech, are contemplating achieving the same goal by requiring researchers to retain copyright over their papers and licensing them to publishers (see *Nature* 396, 293; 1998).

One-stop shopping

Whatever happens, a shake-out of the scholarly publishing market seems inevitable. In his book, *Information Rules: A Strategic Guide to the Network Economy* (McGraw-Hill, 1998), Hal Varian, an economist at the University of California at Berkeley, argues that the economics of the Internet renders unstable the traditional oligopolies achieved by economies of scale. Instead, the Internet tends to be populated by companies with temporarily dominant positions, which can be usurped almost overnight by competitors with better technologies or more attractive features.

The prospect of a few web science publishers roughly equivalent to the online bookseller Amazon.com, offering one site from which you could find the full text of articles from every journal at the click of a mouse, is not so far off.

The promise of 'a library on the desktop' has finally begun to become a reality over the past 12 months, and user demand for the extra capabilities of electronic journals is driving libraries inexorably towards an electronic future. "We are in a transition phase, but electronics will come to dominate because it has all the features and values," says Mike Stout, head of electronic journals at Oxford University Press.



Power to the scientists: HighWire assists learned societies and universities to publish on the web.

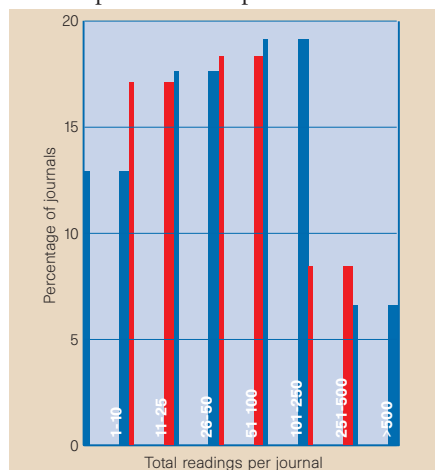
Demand is also being created for sophisticated new products that bear little resemblance to traditional journals. Scientists still strive to publish in brand name journals with the highest impact. But, as consumers, they tend to consult regularly only a handful of electronic journals — the ‘musts’ for their field — and some of the leading journals such as *Nature* and *Science*.

For the rest, individual titles seem less important than the scope for searching and browsing. Using facilities such as citation tracking, researchers are in effect increasingly compiling personal journals for the topics or authors of interest.

This market pressure for ‘one-stop shopping’ at a handful of sites is leading to a proliferation of intermediaries, or ‘aggregators’, which sell packages of journals and other information, and whose ambition is to become the first port of call. The competitive edge of these companies — most of which are American or British — is coming to depend on their ability to muster a critical mass of attractive information accessible through a single powerful and user-friendly interface.

The New York-based company Ovid Technologies, for example, has established itself as a leading provider of electronic material in the biomedical sciences, largely on the basis of sophisticated search software. It provides access to more than 300 journals, and is increasing its portfolio by 30 a month. Ovid was purchased last year by Wolters Kluwer for \$200 million.

Competition at the broadest level is focused on becoming the main gateway to the primary literature. The likely winners are the large abstracting and indexing services, such as the not-for-profit On-line Computer Library Center (OCLC), based in Dublin, Ohio, whose 20 million bibliographic records give it a huge competitive advantage. The Philadelphia-based Institute of Scientific Information’s (ISI) database covers 8,000 primary journals, while databases such as Medline and Embase hold millions of citations for particular disciplines.



Shelf lives: annual consultations of journals in a typical library. Less-read titles could disappear.

Preserving papers for posterity

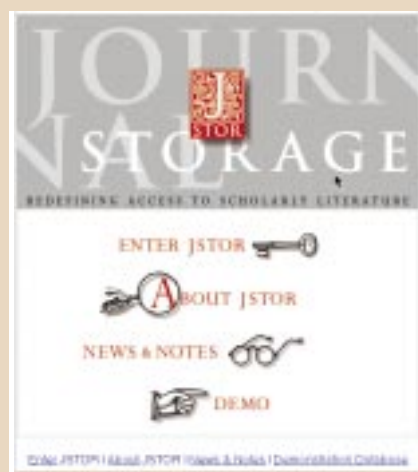
Paper, for all its drawbacks, has one big advantage over electronic media — it can last for thousands of years. The main obstacle to abandoning print journals is the worry that millions of ‘digital objects’ might be unreadable in just a few years because of hardware and software obsolescence.

The problems are mind-boggling. Compact discs only a few years old are often unreadable by new computer systems, and software codes are changing rapidly. The hyperlinked digital content of web material poses even greater problems. How can working links be maintained between text, graphics, audio and video — and who should be responsible?

Technological fixes are possible. Computers could cheaply and quickly migrate one form of software to another across entire systems. Upgrading material to run on the latest equipment is also mostly a matter of management.

But Yola de Lusenet, executive secretary of the European Commission on Preservation and Access, warns against “an unbounded belief in technological solutions”. He advocates greater attention to developing industry standards, to avoid “zillions of different formats”. The ‘millennium bug’ problem has not inspired users’ confidence in the foresight of the computer industry.

A belt and braces approach is being taken by JSTOR — for ‘Journal Storage’ — a non-



Screen saver: JSTOR is archiving back issues of print journals electronically.

profit US body set up in 1995 to archive scientific journals electronically. It scans back issues and generates both printable bitmap images of pages faithful to the original but not searchable, and text files produced using optical character recognition software, which allow searching but may contain errors. It prefers these neutral formats because they should be easier to convert to future formats.

“No single ‘best’ solution to the archiving problem will emerge,” predicts Deanne Marcum, president of the US Council on Library and Information Resources, a body

Until recently, the interest in such services was limited. Scientists could go from a citation to an abstract, but that was about it. But a flurry of deals with publishers over the past year is turning into reality the promise of being able to move rapidly and seamlessly from a citation to full text.

OCLC’s First Search is used by almost all US research libraries, for example, while ISI’s deals with publishers give users access to the full text of more than 100 journals from its web platform, the Web of Science.

In return, collaborating publishers such as HighWire Press can link citations in their journals to ‘Related Records’, an ISI feature that finds articles that have a cited reference in common with the original source article. Similarly, Britain’s Institute of Physics and other publishers are making reciprocal deals to give each other links to their journals.

Web-based communities

But central databases are just one of the many competing and complementary models emerging on the web. An equally pronounced trend away from centralization is also occurring, with a boom in sites tailored to individual communities, such as Biomednet in biology, TipTop in physics and ChemWeb in chem-

istry. Biomednet now has 300,000 registered users, including 3,000 libraries.

Many of these sites include abstracting and indexing services. The Institute of Physics offers the INSPEC physics database, for example, and Biomednet offers an enhanced version of Medline. But their unique selling point is their ‘club’ atmosphere, providing news, features, directories of laboratories and equipment suppliers, and discussion groups — not journals, but fully-fledged electronic magazines.

Taking this concept a step further, many publishers are making ‘knowledge environments’ their top priority — the bringing of a critical mass of resources, services and technology to bear on individual topics. HighWire and the American Association for the Advancement of Science expect to release a series of ‘knowledge environments’ this year. “We think there are good reasons to create deep but narrowly construed digital libraries,” says HighWire’s Michael Keller.

This is a profound shift away from the traditional journal system. Users and librarians increasingly view titles as merely part of hyperlinked ‘content databases’ made up of constellations of journals. “We are trying to get away from single journals and think

that includes the US Commission on Preservation and Access. "Only practical experience will help us make some of the hard decisions that lie ahead."

One of the biggest questions is who should be responsible. Libraries have traditionally archived journals, but specialized agencies have dominated archiving of video and audio products. Centralized national depositories may similarly be more appropriate for digital material, says Don Waters, director of the US Digital Library Federation.

But Waters believes that many of the issues are best dealt with locally by the "interested communities". Kevin Guthrie, president of JSTOR, agrees: "Individual organizations committed to archiving their material are best placed to solving their local [logistical and technical] problems."

The American Physical Society has made archiving a priority, and has digitized issues of its *Physical Review* back to 1985 at a cost of \$750,000. "I scanned in volume 1 from 1893 this morning, so now all we need to do is fill in the gap," quips editor in chief Martin Blume. Copies are kept at several libraries to guard against mishaps.

A major concern is whether commercial publishers, which increasingly archive their material, will maintain a long-term commitment. Librarians are worried that publishers will abandon costly and little used archives. Given this risk, and the

uncertainty created by possible bankruptcies or buyouts, librarians are keen that archiving should be a public service.

"Publishers may find it commercially sensible to off-load this eventually to a national archive where access could be managed according to either the terms of a licence from the publisher or under statutory requirements for legal deposit," says Graham Cornish, from the British Library. Ultimately "legislation will be needed to regulate archiving," he adds.

Questions of who is responsible for archiving are likely to be easier to resolve in Europe than in the United States, points out Marcum. Europe has national agencies for archiving information, whereas the closest US equivalent, the Library of Congress, has no national mandate for such activity.

For its part, JSTOR expects to hold complete runs of more than 100 titles in 10 to 15 disciplines by the end of this year. But Ann Okerson, associate librarian at Yale University, argues that JSTOR is unlikely to be a panacea because it is archiving only the leading journals. "These would be the last to disappear," she points out.

Guthrie accepts this criticism, but says that as a non-profit organization it had to start with journals that were likely to allow it to recover its costs through fees charged to libraries for access. "It would be great to archive more obscure titles but hard to find people willing to pay," says Guthrie.

more in terms of developing knowledge environments that integrate a number of relevant sources," says Mike Stout of Oxford University Press.

Who needs journals?

This view of the literature as one vast interwoven content database is in turn leading many to question the utility of the traditional compartmentalization of this information with thousands of journal titles.

The 'serials crisis' has provoked widespread dissatisfaction with the cost of the journal system. Some economists argue that many obscure high-priced, low-circulation journals should exist only electronically if at all, on the grounds that the costs of production in print are prohibitive.

In a typical library, half the journals are consulted no more than 50 times annually, and only 15 per cent more than 250 times, according to Carol Tenopir, an information scientist at the University of Tennessee, Knoxville, and Donald King, a consultant based in Ann Arbor, Michigan (see graph opposite).

But the cost of quality typesetting accounts for \$500 of the estimated average \$2,000 cost of producing a 20-page article,

according to Berkeley economist Hal Varian. This works out at \$5 per person if 100 people read the article, and \$50 if only ten read it.

The conclusion he draws is that, for many journals, these costs are simply not worth it and production standards should be dropped, with typesetting done by the authors. High-quality typesetting would be reserved for higher-circulation journals.

The Los Alamos e-print archives, to which authors submit formatted articles, is a working example of Varian's proposal. Andrew Odlyzko, an expert on the economics of electronic publishing, estimates that each e-print costs between \$5 and \$75 to produce, even taking into account the subsidy of \$1 million over three years that the archives receive for software development from the US National Science Foundation.

Odlyzko predicts that, while high-circulation journals will exist in print and electronic versions for the foreseeable future, the low-circulation, high-cost journals that make up the bulk of the scholarly publishing market will in future exist only electronically.

These arguments are endorsed by Michael Keller at HighWire Press. "I think that the solution to the 'journals crisis' (see page 195) lies in creating a situation where

the journals of very low circulation are entirely supplanted by electronic editions, and where the journals have got to be produced by and for the practitioners themselves."

Information overload

At a wider level, there seems to be growing acknowledgement that the main role of journals in future will be to provide research papers with a guarantee of quality and added editorial value — in terms of making the science more readable, and placing it within a wider perspective for example — while their traditional role as a distribution outlet will become less important.



Ginsparg: pioneer of physics e-prints.



Blume: 'publishers add value to papers'.

The Los Alamos physics archives created by Paul Ginsparg in 1991 now receive some 25,000 new electronic papers annually and have become the primary means of communication for physicists, tens of thousands of whom connect daily. This reality has been acknowledged by traditional journals, and the American Physical Society and several other publishers which now link their journals to the archives.

"We provide an imprimatur and a collection of articles certified as worthy of attention," says Martin Blume, editor in chief of the American Physical Society. "The individual article is not the valuable element, it is [added value put into creating] the collection."

Ironically, electronic publishing is fuelling demand for islands of filtered information in the ocean of information now available. Information overload is frequently cited by scientists as the biggest problem they face in using the web.

The Los Alamos archives have acknowledged this and have arranged with journals such as the *Journal of Artificial Intelligence Research* and the *Journal of Applied and Theoretical Mathematics* to provide peer-reviewed overlays to the unfiltered archives.

Moves are afoot to create a similar e-print archive for the biomedical sciences (see *Nature* 397, 91; 1999). The idea is being championed by Patrick Brown, a researcher at Stanford University School of Medicine, and David Lipman, director of the US National Center for Biotechnology Information, which runs PubMed and GenBank.

It is too soon to predict the prospects for Brown's initiative. But the previously antagonistic attitude of many publishers to the electronic prior publication of articles has

become more liberal over the past two years. Publishers that refuse a priori to publish articles that have been posted on e-print servers — such as the *New England Journal of Medicine* — appear increasingly isolated.

In contrast, over the past two years, *Nature*, the *Journal of Neuroscience* and several other journals have stated that posting on e-print servers does not a priori constitute prior publication, but is rather a legitimate means of communication between researchers (see *Nature* 390, 427; 1997).

The *British Medical Journal* joined their ranks this month. In an editorial, Richard Smith, its editor, argues that journals have nothing to fear from e-print servers. “Strong publication is associated with prestige, credibility, reliability, wide availability, news coverage and permanence... [scientists] want to publish both on e-print servers and in peer-reviewed journals. It’s not either/or but both.”

Brown wrote this month to major journal publishers asking them to publish “an explicit policy statement that distribution of a preprint, by means of a public electronic

preprint server or Internet site, will not influence the decision of your journal to publish a paper”. This would remove one of the major deterrents to wider use of e-print archives.

Virtual peer review

Several journals, including the *BMJ*, are experimenting with making manuscripts available on the web before they have been peer reviewed, and then subjecting them to open, online peer review.

A pilot test on one article prompted a large response from readers. Tony Delamothe, deputy editor, admits that the opinions expressed were of variable quality, but believes that they nonetheless allowed conclusions to be drawn as to whether the paper should be accepted for publication.

The stakes are high, points out Delamothe, given that, in contrast to physics, a change in publication practices could have public health consequences as information about potential treatments would be made public before being validated scientifically. But the journal is optimistic that labelling

non-peer-reviewed material may be sufficient to prevent abuse.

The journal intends to carry out further controlled experiments before changing its editorial policies. It wants especially to establish whether naming referees might affect the quality of reviewing — young referees might refrain from publicly criticizing their elders, for example, for fear of retaliation.

One idea that the *BMJ* may consider is a hybrid peer-review model that combines open online peer review and commissioned reviews. This strategy is being pursued by *Electronic Transactions in Artificial Intelligence*. It differs from conventional journals in that review and acceptance take place after the article has been published online.

The system is yielding better quality papers than conventional reviewing, claims Erik Sandewall, the journal’s editor. He adds that open online reviewing “broadens the concept of scientific publication so that the feedback and quality-control processes become integrated with the author-to-reader communication instead of being separated from it as at present”.

A less remarkable change, but one welcomed by researchers, is the practice of posting papers on the web upon acceptance, often many weeks before their appearance in print. “This is key for me; it is a tremendously good thing,” says Gregory Fu, a chemist at the Massachusetts Institute of Technology.

Electronic publishing is stimulating other innovations in the submission process. The entire editorial procedure of the *Journal of High Energy Physics* is managed by a software robot, which scans papers submitted by e-mail and assigns them to referees on the basis of key words. Authors, editors and referees have real-time access to papers throughout the editorial process.

Unfortunately for publishers and librarians, nobody has invented a software robot that can design winning strategies amid the Brownian motion of the electronic publishing business. Martin Blume of the American Physical Society sums up what many consider will be the only realistic web strategy for some time: “Experiment as much as possible, and be as fast on our feet as we can.” □

Roll over Gutenberg

Many scientists remain strongly attached to the ‘look and feel’ of the printed page, and doggedly continue to download and print web documents, rather than reading them on screen. In doing so, they are contributing to perhaps the largest, though inconspicuous, paradigm shift yet to have been brought about by electronic publishing — the shift from centralized printing to electronic distribution and local printing.

So much for visions of the paperless office. Over the past five years paper consumption has jumped 13 per cent in the United States, with 1,000 billion pages pouring out of computer printers annually.

Some publishers, such as the American Physical Society, are considering whether it might not be cheaper to stop printing low-circulation journals, and just let libraries, or whoever wants paper copies, download files and print whole issues themselves.

The US company Presspoint is already exploiting this idea to print short runs of foreign newspapers in hotels and airports. The digital printing presses required, which skip the conventional preparation of individual typeset pages on film, are becoming cheaper and more widely available. For short print runs they are as or more economical than traditional printing — and quality is the same.

But why do researchers download and print? One answer is that, although even the best monitors may look sharp, they are fuzzy, and their resolution is well below the 200 pixels per inch or so that would make reading as comfortable as on paper. Reading on screen is slower and more tiring.

The first of an expected wave of digital reading devices — or electronic books — Nuvomedia’s Rocket eBook, went on sale before Christmas at \$499. Each can hold the equivalent of a dozen novels, and offers touch sensitive, high-resolution screens.

The content of e-books will initially be restricted to special encrypted book titles downloaded from the web. But Nuvomedia, which has agreements with several major publishers, is looking at other markets. “We definitely have plans to pursue journals,” says Nuvomedia’s Robert Carter.

If Nuvomedia’s vision of scientists carrying their personal libraries around with them seems far fetched, that of E Ink, a company born at the Massachusetts Institute of Technology, appears almost science fiction. E Ink has invented an electrophoretic ink of microscopic coloured capsules that change colour when a tiny electric current is passed through them (see *Nature* 394, 253–255; 1998). Coat the ink onto paper, plug the sheet into a computer, and the sheet can produce high resolution images — black and white at present — that stay when the current is switched off.

Russ Wilcox, E Inks’ vice-president, claims the ink could be used to develop screens with four times the resolution of existing screens. The company plans to create paper books that could display any electronic text. Might researchers soon be able to download their copy of *Nature* and carry it with them on the train? “Absolutely,” says Wilcox. “Electronic ink’s light weight and low power draw make it ideal for such portable applications.”



Close the South–North knowledge gap

Sir—The rising cost of journals and difficulties associated with hybrid journals limit access to knowledge by scientists in poorer countries¹. The economic and technical difficulties that contribute to the North to South knowledge gap will not be resolved until alternative mechanisms for the distribution of information are developed and scientific societies take steps to revise the present tradition. Until then, scientists in developing nations will continue to be disenfranchised.

Although the North to South gap is widely acknowledged, the gap from the South to the North is less appreciated. Yet this deprives the global scientific community of much essential information from developing countries. It is caused by problems faced by publishers in these countries in meeting the costs of printing and distributing their peer-reviewed journals. Scientists in such regions have difficulty publishing in high profile journals. As Richard Horton, editor of *The Lancet*, has said, “The invisibility to which mainstream science publishing condemns much Third World research thwarts the efforts of poor countries to strengthen their

journals — and the quality of research — in regions that most need them”.

Fortunately, electronic publishing can resolve many of these problems (see Briefing, page 195). The feasibility of this has been shown by organizations such as the Electronic Publishing Trust for Development (EPT)² and workshops organized by the British Council³. The EPT has facilitated the online publication of 16 peer-reviewed bioscience journals in Africa, Asia, Central and South America. With a small investment, publishers can readily learn to prepare their publications in web-compatible format and benefit from the increased visibility. The independence so gained allows developing countries to establish their own distribution sites, so strengthening their science base.

Thanks to online journals much previously unknown research now forms part of the international knowledge base. The heightened awareness that electronic distribution provides leads to renewed enthusiasm for publishing in local journals, and the sense of isolation often felt by the scientific community begins to diminish.

The gap from North to South will take

time to close as new mechanisms are developed and attitudes change. The gap from South to North can be closed more swiftly since the technology is easy and low cost and, importantly, access to the Internet is not immediately essential if partnerships can be made with non-profit facilitating organizations and scientific societies.

Vanderlei Canhos*, **Leslie Chan†**, **Frances Giaquinto‡**, **Barbara Kirsop§**, **Anthony O'Donnell¶**, **Judy Ugonna#**

*University of Campinas/Base de Dados Tropical, R. Latino Coelho, 1301 Parq Taquaral, 13087-1010 Campinas, SP, Brazil

†Centre for Instructional Technology Development, Bladen Library, University of Toronto, Ontario, M1C 1A4, Canada

‡BioDiversity, Brunel Science Park, Kingston Lane, Uxbridge, UB8 3PQ, UK

§EPT Secretariat, Stainfield House, Stainfield, Bourne, Lincs PE10 0RS, UK

¶Department of Agriculture, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK

#British Council, Medlock Street, Manchester M15 4AA, UK

1. Annapoorani, B. et al. *Nature* **395**, 739–740 (1998).

2. <http://dspace.dial.pipex.com/biolinel/>

3. <http://www.britcoun.org>

Luddites must not block progress in genetics

Sir—You report Monsanto's indecision about promoting its 'terminator' technology for sterilizing crop seed (*Nature* **396**, 503; 1998). Terminator is becoming a classic case of 'transnational Luddism'.

Monsanto will profit most if the technology spreads to developing countries, reducing their need to import grain. Monsanto's self-interest in exporting seed technology is the opposite to the economic interests of North American farmers who produce most of the world's grain exports. Opponents of technology transfer — including the 'prairie' non-governmental organization (NGO), the Rural Advancement Foundation International mentioned in your report — sensibly try to protect American grain exports by lobbying worldwide against Monsanto. It seems that national economic interests are best met by developing 'terminator' and, more widely, the technology of genetically modified organisms (GMOs), yourself and then persuading competitors overseas not to use them by exporting Luddite anti-technology scaremongering.

Why all the fuss now about sterile seed?

For several millennia farmers have propagated bananas clonally, producing sterile triploids — 'terminated' bananas. No-one can steal farmer-varieties by planting banana seed: banana farmers are well in advance of Monsanto in protecting intellectual property.

As for GMOs, for more than 5,000 years wheat has been a genetic monster with entire 'alien' complements of genes from three species. This wide hybridization allowed wheat to spread to just about everywhere on the agricultural frontier. If plant breeders tried to repeat this miracle now to feed developing countries there would be an outcry from Luddite NGOs and conservationists: research would be halted, and there would be no food for more than a billion people.

Farmers in developing countries do not need the NGO mixture of paternalism and export protectionism, disguised as 'in the farmers' interests'. It is patronizing to claim that farmers will be 'hooked' on sterile seed technology: they will make rational decisions in their own interests. And one suspects that Third World farmers do not relish the role assigned to them by these NGOs of museum-keepers of obsolete varieties for our plant breeding needs.

David Wood

13 Herons Quay, Milnthorpe, Cumbria LA7 7HW, UK

Computers and the copyright conundrum

Sir—The News report entitled “Personal computers spur drive to keep control over copyright” appears to misapprehend important aspects of copyright law (*Nature* **396**, 293; 1998).

It describes universities and publishers as battling over the copyright to contributions by academic authors and states: “One of the main issues of concern is how to maintain the peer-review certification process. . . while allowing electronic publishing to disseminate scholarly work more widely.” But the connection between peer review and copyright is far from obvious. Peer review does not depend in any important way on who owns the copyright. Under long-standing tradition, the mere submission of a manuscript for publication in a peer-reviewed journal grants an implied licence to make the adjustments called for by the peer-review process. In any event, the publishers could demand such a licence expressly, without demanding a transfer of the entire copyright to the article.

The real issue is: will electronic publishing (or other forms of distribution) by the author or university reduce demand

for the journal? Scientific publishers cannot be expected to absorb the costs of running articles through peer review and printing them if they will lose most of their sales because readers get their copies direct from the authors (who will note proudly on their manuscripts that the articles have been peer reviewed and accepted by X Journal).

This issue arises in theory even if the journals do get an assignment of the copyright, if authors as a group are successful at negotiating licences to distribute their work. Even without a licence, the author may be free after transfer of the copyright to make at least some kinds of distribution as a 'fair use'.

In the United States, whether the scholar or the university owns the copyright has not been firmly decided. A straightforward reading of the Copyright Act leads to the conclusion that the university owns the copyright (subject to negotiation with the members of the faculty) as a work made for hire. But at least two important judicial decisions have found a 'professor's exception' to the work-for-hire doctrine, based on long tradition at universities.

Dennis S. Karjala

Arizona State University, Tempe,
Arizona 85287-7906, USA

Who pays what in drug development

Sir — Henry Miller¹ repeats the much quoted claim that "bringing a drug to market in the United States now costs more than \$500 million," adding that this is "by far the highest price tag in the world".

Miller's letter gives the misleading impression that drug companies spend hundreds of millions of dollars on the tests needed for US Food and Drug Administration marketing approval. The costs of drug development are not trivial, but the figures Miller cites are based mostly on estimates of the costs of preclinical research, rather than FDA regulatory burdens. The most detailed study of the costs of clinical trials was a 1991 *Journal of Health Economics* paper². In 1997 prices, the average out-of-pocket costs of clinical trials needed for FDA approval were \$25 million. Adjusted for risk, the 'per approval' cost of clinical trials was \$56 million.

The \$500 million figure quoted by Miller and others adjusts these costs somewhat higher to include 'capital costs' for financing trials, but also and most importantly the cost of preclinical research, which accounts for 70 to 80 per cent of the total cost of drug development in some studies.

Moreover, it is often governments rather than the drug companies that pay for

clinical and preclinical research. For example, according to US tax returns, from 1983 to 1993 the pharmaceutical industry reported expenditure of only \$213 million on clinical trials for orphan drug development. This was about \$2.3 million for each of the FDA's 93 orphan drug approvals during the period.

James Love

Consumer Project on Technology, PO Box 19367,
Washington, DC 20036, USA

1. Miller, H. *Nature* **395**, 835 (1998).
2. DiMasi, J. A., Hansen, R. W., Grabowski, H. G. & Lasagna, L. *J. Health Econ.* **10**, 107–142 (1991).

Guard your knowledge... and reap the rewards

Sir — The British government's desire to put "the commercialization of scientific knowledge at the heart of its industrial policy" is timely (*Nature* **396**, 714–715; 1998). The ideal is a seamless integration between public and privately funded research with scientists having interests in both sectors.

However, what constitutes knowledge generated from academic, publicly funded research? Knowledge is any privileged information which can generate or add value to intellectual property and, when exploited, can be sold for profit. A conversation over coffee or a discussion at a poster cannot be valued but nevertheless is the transfer of knowledge. So is the perusal of grant applications and papers submitted for publication. At present the exploitation of this knowledge is unregulated.

Without safeguards and procedures to ensure that knowledge, however defined, is properly valued, funding agencies will lose rewards to which they should be entitled.

David Grant

Leukaemia Research Fund,
43 Great Ormond Street, London WC1N 3JJ, UK

German university reforms misguided

Sir — Your editorial "When payment by results is a sensible approach" was ill-informed and superficial (*Nature* **396**, 393; 1998). You highlighted suggestions from Germany's University Rectors' Conference (HRK) that new performance-dependent criteria should be applied to universities to increase efficiency and cut costs: that the pay of senior academics should be performance dependent, and that lifetime tenure should be abolished.

These suggestions would exacerbate the

problems. First, it is true that German professors receive a sum of money for running costs which is not dependent on their research performance. However, often this sum barely covers the costs of teaching and departmental maintenance which are aggravated by service costs for large equipment as required by law. Running a modern research programme on this money is generally out of the question.

Second, a performance-dependent scale that is partially indexed by the amount of external funding is a questionable criterion. Within Germany there is a large pool of industrial money which is extensively used by academics. However, many of these collaborative grants are scientifically mundane and unchallenging. The criteria for success in winning such a grant are fundamentally different to those for public-sector grants. At present the only funding source that guarantees scientific excellence is the Deutsche Forschungsgemeinschaft, which is coming increasingly under financial pressure. So, judging performance solely by the amount of external funding would drag German science in the direction of mediocrity.

Another major difficulty is the assignment of performance indexing to a body within the university. Major decisions in German universities are often highly political and subjective. At present, this situation is usually tolerable as the personal status of the individual is not open to influence. The HRK suggestions seek to change this. I dread to think of the incestuous consequences this would bring.

The suggestion that tenure should be abolished to improve efficiency is at best surprising. All other European countries award tenure to academics on the assumption that, despite possible negative consequences, it is essential to ensure long-term planning for scientific programmes.

Many see the HRK developments as a political manoeuvre on the part of the technical high schools (*Fachhochschulen*), which are essentially teaching establishments, to attain full university status without having to fulfil the present requirements for research excellence. This has been a point of contention for years, as salaries in the *Fachhochschulen* are slightly lower than in the universities but would become equal under the HRK recommendations because the university professorial salary would be lowered.

The real problems facing university academics are that research groups are often too large and run on an imperialistic basis, and that funding opportunities for young scientists to organize independent research groups are too limited.

Robin Ghosh

Institute of Biology, University of Stuttgart,
Pfaffenwaldring 57, D-70550 Stuttgart, Germany

Prospecting for Jurassic slabs

Mark A. Richards

Increasingly precise seismic imaging of the Earth's deep interior continues to sharpen our understanding of mantle convection and plate tectonics. One such study has traced events down to 2,800 km beneath Earth's surface, and back to 200 million years ago.

Seismic tomography, named in analogy to medical imaging, is a child of the early 1980s now grown to vigorous adulthood. It is the technique involving analysis of seismic waves that travel through the Earth and are affected in various ways by what they pass through *en route* from source

to sensor. Among other things, seismic tomography has shown that flow in the Earth's mantle is organized into concentrated downwelling and upwelling structures. High-velocity anomalies, where seismic waves speed up, are comparatively common and correspond to regions where cold litho-

spheric plates have sunk into the mantle at the convergent margins of tectonic plates¹ (subduction zones) (Fig. 1a). Concentrated low-velocity (hot) structures are less prevalent, the largest lying beneath southern Africa and causing considerable uplift of the African continent².

On page 246 of this issue³, Van der Voo *et al.* advance the geological interpretation of seismic tomography by recognizing the likely connection between a major high-velocity anomaly lying 1,500–2,800 km beneath Siberia and the subduction of an oceanic plate during Jurassic time, about 150–200 million years ago. This discovery is significant in several respects — it is the oldest slab of subducted lithosphere individually recognized in the deep mantle; it supports the idea that some slabs sink essentially to the bottom of the mantle¹; and it calibrates the timescale for sinking. Furthermore, this work shows that seismic tomography is a powerful tool for palaeogeography, providing a fresh source of information about the history of global plate motions and the processes of continental growth.

This level of maturity in understanding mantle dynamics stems from continuous improvements in both seismic imaging and its geodynamic interpretation. In the mid-1980s it was discovered that large-scale (10,000–20,000-km) horizontal variations in seismic velocity represent mantle density anomalies that give rise to the shape of the Earth's gravity field, or geoid⁴. This large-scale structure is a consequence of where cold lithospheric plates have sunk into the mantle during the past 200 million years of Earth history⁵ (Fig. 1b) — that is, most of Cenozoic and Mesozoic time. But subducted slabs in the deep mantle are only a few hundred kilometres thick and could not be well resolved in early tomographic studies. Now they can be, allowing detailed geological interpretation of the history of convergent plate margins in particular.

Seismic tomographers employ many types of seismic vibration to image the deep Earth. But the most compelling images have been produced using higher-frequency 'body' waves travelling from earthquake sources to the worldwide network of seismometers. The speeding-up or slowing-down of a seismic ray as it passes through a high- or low-velocity region perturbs its arrival time. If the locations of the earthquakes are well known, thousands of such measured arrival times (and waveforms) can be mathematically 'inverted' to provide a three-dimensional model of seismic velocity structure in the mantle.

This kind of analysis is highly complicated, requiring powerful computers and huge amounts of seismic data. In 1994, however, a breakthrough came in a study, published by Grand⁶, in which body shear-waves were used to image the extinct Farallon plate,

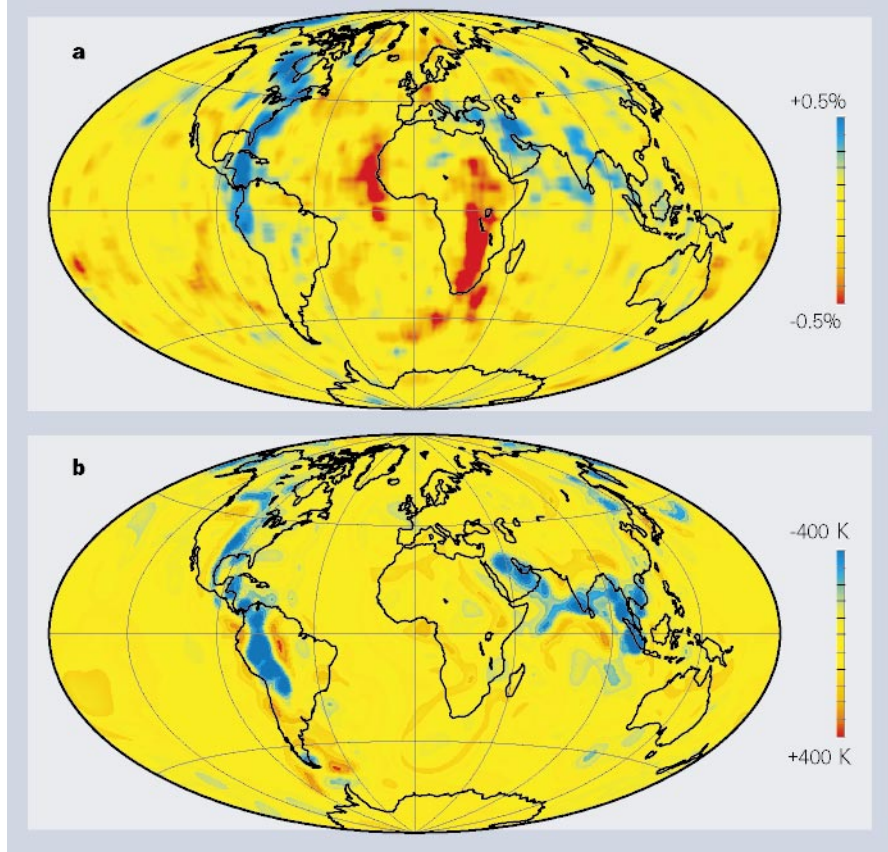


Figure 1 Comparison of models of mantle structure at 1,500 km depth. a, Shear-wave tomography model¹ (velocity variations in percent; blue is high velocity, red is low velocity). b, Geodynamic, or subduction, model¹⁴ (temperature variations in kelvin; blue is low temperature, red is high temperature). The geodynamic model accounts for thermal convection beneath surface plates whose motions are derived from plate tectonic reconstructions for the past 120 million years. The two models are similar in that structure is dominated by cold, high-velocity anomalies corresponding to subduction zones. But they differ in two important respects. First, concentrated hot (upwelling) structures are absent in the geodynamic model. Second, the positions of subducted slabs in the geodynamic model correspond only roughly to those observed seismically: for example, the seismic signature of the Farallon plate beneath North America lies about 1,000–1,500 km east of its location in the geodynamic model. Also, the geodynamic model has too little convergence between Europe and Africa, and so fails to explain the high-velocity anomaly beneath the eastern Mediterranean region. The top of the sub-Siberian high-velocity anomaly, the subject of the new study by Van der Voo *et al.*³, is seen faintly with shear-wave tomography, but is completely absent in the subduction model, which does not include the Mongol–Okhotsk subduction.

sinking from 700-km to 2,800-km depth beneath North and South America (Fig. 1a). Subsequently, using high-resolution compressional-wave tomography, Van der Hilst *et al.*⁷ further demonstrated the prevalence of sheet-like high-velocity structures beneath past and present subduction zones, some extending to the boundary between Earth's liquid outer core and rocky mantle. The difference between shear and compressional waves need not concern us here. Rather, the point is that these entirely independent data sets delivered spectacular agreement between images in areas where the quality of seismic ray coverage was similar¹.

Van der Voo *et al.*³ now take 'seismoslabology' a step further in relating a hypothesized case of continental accretion in Jurassic time to a hitherto unexplained high-velocity seismic anomaly beneath Siberia. Large continents such as Eurasia are formed by collisions of smaller continental fragments, preceded by subduction of the intervening oceanic plate. In Late Jurassic time, as a block of continental fragments (including much of present-day Mongolia, China and Southeast Asia) moved towards Siberia from the southeast, the Mongol–Okhotsk Ocean basin closed⁸ (see map on page 249). The Mongol–Okhotsk oceanic plate was consumed, and the Verkhoysansk Mountains were pushed up, as the continental blocks collided.

Subduction ended by the Early Cretaceous, or about 135 million years ago, so that a large slab of oceanic lithosphere became detached from the overlying lithosphere as it sank into the deep mantle. Improved tomographic resolution of sub-Eurasian structure⁹ reveals a large, high-velocity (cold) anomaly sinking in the mantle just below the hypothesized site of the Mongol–Okhotsk subduction, with the top of the anomaly being at about 1,500 km depth. Coupled with the likely date at which subduction stopped, this implies that the slab has sunk into the deep mantle at about 1 cm yr⁻¹ — slower than, but roughly in accord with, previous estimates^{6,10}. The anomaly extends to the core–mantle boundary, where it joins a much broader high-velocity structure that underlies much of eastern Eurasia and also connects with active subduction zones in the western Pacific.

The Mongol–Okhotsk slab seems to have undergone little horizontal displacement since Jurassic time relative to the present-day 'suture zone' of the continental collision. Even more remarkably (and perhaps coincidentally), it preserves the 'Z' shape, in map view, of the old subduction zone along the margin of the Siberian block. Although other possible explanations cannot be ruled out entirely, that of Van der Voo *et al.*³ looks to be plausible in all respects.

The interpretation of the sub-Siberian

high-velocity anomaly in terms of Mongol–Okhotsk subduction supports models of supercontinent aggregation that invoke subducted slabs as the driving mechanism for convergence between continental blocks¹¹. Straightforward modelling of the oceanic lithosphere as the upper thermal boundary layer for mantle convection shows that slabs should survive as concentrated thermal anomalies during their 100–200-million-year journey to the bottom of the mantle, a view supported by the interpretation of Van der Voo *et al.*

The case is not yet completely closed, however. Some slabs appear to be inhibited as they encounter the mantle transition zone (400–800-km depth), where a series of mineral phase transitions induced by increasing pressure may cause complicated buoyancy and rheological effects^{7,12}. Another concern is that the sub-Siberian anomaly appears to maintain a faint high-velocity connection all the way to the surface, even though subduction ceased by the Early Cretaceous. Although this upper mantle extension is poorly resolved by the seismic ray coverage used, better imaging of the sub-Siberian region should aid in refining interpretation of the large, deep-mantle anomaly.

Seismic tomography is providing similar insights into the fates of other subducted slabs. For example, the old Farallon plate is found^{1,6,7} about 1,000–1,500 km to the east of where a simple subduction history model might place it (compare Figs 1a and b). This kind of information is important, because

subducted slabs are probably the largest dynamic source of mantle buoyancy — which, among other things, controls the height of continents and the stratigraphic record of sea level¹³. As illustrated by the results discussed here, the new generation of seismic tomography tools provides geologists and geophysicists with tantalizing food for thought. It is allowing them to refine models for past plate motions and mantle convection, and so arrive at a more complete picture of Earth history during Cenozoic and Mesozoic time. □

Mark A. Richards is in the Department of Geology and Geophysics, University of California, Berkeley, California 94720, USA.

e-mail: markr@seismo.berkeley.edu

- Grand, S. P., Van der Hilst, R. D. & Widiyantoro, S. *GSA Today* **7**, 1–7 (1994).
- Lithgow-Bertelloni, C. & Silver, P. G. *Nature* **395**, 269–272 (1998).
- Van der Voo, R., Spakman, W. & Bijwaard, H. *Nature* **397**, 246–249 (1999).
- Hager, B. H., Clayton, R. W., Richards, M. A., Dziewonski, A. M. & Comer, R. P. *Nature* **313**, 541–545 (1985).
- Richards, M. A. & Engbreton, D. C. *Nature* **355**, 437–440 (1992).
- Grand, S. P. *J. Geophys. Res.* **99**, 11591–11621 (1994).
- Van der Hilst, R. D., Widiyantoro, S. & Engdahl, R. *Nature* **386**, 578–584 (1997).
- Ziegler, A. M. *et al.* in *The Tectonic Evolution of Asia* (eds Yin, A. & Harrison, M.) 371–400 (Cambridge Univ. Press, 1996).
- Bijwaard, H., Spakman, W. & Engdahl, R. *J. Geophys. Res.* (in the press).
- Ricard, Y., Richards, M. A., Lithgow-Bertelloni, C. & Le Stunff, Y. *J. Geophys. Res.* **98**, 21895–21909 (1993).
- Gurnis, M. *Nature* **332**, 695–699 (1988).
- Fukao, Y., Obayashi, M., Inoue, H. & Nenbai, M. *J. Geophys. Res.* **97**, 4809–4822 (1992).
- Gurnis, M. *Nature* **364**, 589–593 (1993).
- Bunge, H.-P. *et al.* *Science* **280**, 91–95 (1998).

Protein transport

Molecular motors join forces

Manfred Schliwa

The apples you buy at the grocery will usually have completed a long journey, including shipment from the farmer to your home town then delivery to the grocer. Different modes of transport may have been used, such as a train for the long haul and a van for the local dispatch, and the legs of this voyage are coordinated by closely interacting business partners. Similarly, animal cells may use different transport systems to deliver membrane-bound components, such as secretory vesicles, from the site at which they're packaged in the centre of the cell to their destination at its periphery. Long-range movements are carried out by molecular motors ('trains'), using microtubules as their tracks. Often, however, microtubules do not reach the plasma membrane, and a different set of motors ('vans') carry out the last leg. These motors move on actin filaments, so they can traverse the actin-rich cell cortex.

According to a report by Huang *et al.*¹ on page 267 of this issue, there is one big dif-

ference between transporting vesicles and apples — cells combine the two different means of transport in one bifunctional vehicle. The authors provide convincing evidence that a microtubule motor, conventional kinesin, and an actin motor, myosin Va, interact directly (Fig. 1).

Conventional kinesin (the prototype of the growing kinesin superfamily) is believed to be involved in many forms of microtubule-mediated vesicle and organelle transport². Myosin V is one of at least a dozen classes of unconventional (that is, non-filamentous) myosins that move on actin filaments³. Analyses of myosin V mutants in organisms as divergent as yeast and mouse indicate that these myosins are also organelle motors. For example, mice with defects in the *dilute* locus, which encodes myosin Va, have a washed-out ('diluted') coat colour. This is because the transfer of pigment organelles (melanosomes) from melanocytes to the skin cells involved in hair formation is compromised, so melanosome

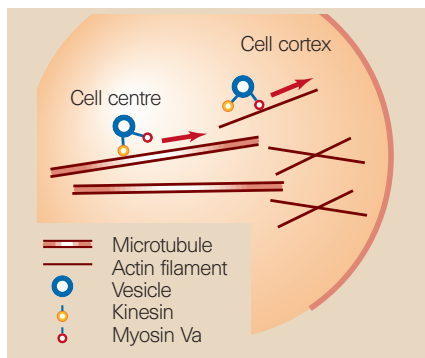


Figure 1 Vesicle transport in the cell by a bifunctional vehicle. Huang *et al.*¹ show that the molecular motors kinesin and myosin, which run on microtubules and actin filaments respectively, interact on the vesicle membrane.

transport is thought⁴ to be mediated by myosin Va.

Huang *et al.*¹ now show that a region in the myosin Va tail, termed the AF-6/cno homology domain, specifically interacts with a roughly 100-amino-acid segment of the conventional kinesin tail. They identified these interacting subdomains using a yeast two-hybrid screen. The interaction is highly specific, and is restricted to the ubiquitous isoform of conventional kinesin — not even the closely related neuronal kinesin interacts. The authors also found that the two motors bind to one another *in vitro* and that they co-immunoprecipitate from mouse brain extracts, making a strong case for a specific interaction *in vitro*. But the evidence to show an association *in vivo* is less revealing, largely because both motors are distributed in a punctate and dispersed fashion in the cell line used, as seen by immunofluorescence microscopy.

The cooperation and interdependence of actin- and microtubule-based cytoskeletal systems has long been known⁵. But the discovery that the motors associated with these two cytoskeletal fibre systems are also coordinated in membrane traffic is more recent. For example, coordination of kinesin's activity with that of the microtubule motor dynein has been shown in a reconstituted *in vitro* system for the bidirectional movement of melanophore pigment granules⁶ and phagosomes⁷. Other studies point to a close interrelationship between actin and microtubule motors in membrane transport. Kuznetsov *et al.*⁸ were the first to suggest that the same vesicle can move on both microtubules and actin filaments. Mitochondria of cultured axons⁹ and pigment granules of melanophores¹⁰ also use both of these tracks. Moreover, in sea urchin eggs, exocytic vesicles required to close wounds in the plasma membrane are recruited by a two-step transport process that involves first a microtubule motor (kinesin) and then a myosin¹¹. So, evidence for cooperation of microtubule- and actin-dependent motors is mounting,

although we don't yet know how the activities of the two motors are coordinated.

The new twist offered by Huang *et al.*¹ is that the motors may interact directly, a possibility not previously considered. Their findings open an exciting direction for future research, and may eventually lead to an explanation for some puzzling observations such as the localization of myosin Va in spindles¹² and centrosomes¹³. However, it remains to be seen whether the principle is a general one. The kinesin involved in amphibian melanosome movements is probably kinesin II, not conventional kinesin. Moreover, the yeast Smy1 (a kinesin-related protein) and Myo2 (a class V myosin¹⁴), which clearly interact genetically, lack the conserved interacting domains identified by Huang and colleagues.

As always, big research challenges lie ahead. How is the kinesin/myosin motor 'bouquet' attached to membranes? Is this a long-term attachment, or are complexes recruited from the cytoplasm when needed? Once attached, how is activity of the motors regulated? Possibly, as in many other cellular processes¹⁵, the answer will be a large protein assembly that coordinates and regulates these activities. The direct interaction of

molecular motors may just be our first glimpse at such a motor-protein machine. If confirmed, the transport of cellular goods — just as the distribution of apples — may be in the hands of closely interacting (molecular) partners. □

Manfred Schliwa is at the Adolf-Butenandt-Institut, Zellbiologie, University of Munich, Schillerstrasse 42, 80336 Munich, Germany.

e-mail: schliwa@bio.med.uni-muenchen.de

- Huang, J.-D. *et al.* *Nature* **397**, 267–270 (1999).
- Lane, J. & Allan, V. *Biochim. Biophys. Acta* **1376**, 27–55 (1998).
- Cope, M. J., Whisstock, J., Rayment, I. & Kendrick-Jones, J. K. *Structure* **4**, 969–987 (1996).
- Wei, Q., Wu, X. & Hammer, J. A. III *J. Musc. Res. Cell Motil.* **18**, 517–527 (1997).
- Schliwa, M. *The Cytoskeleton: An Introductory Survey* (Springer, Vienna, 1986).
- Rogers, S. L., Tint, I. S., Fanapour, P. & Gelfand, V. I. *Proc. Natl Acad. Sci. USA* **94**, 3720–3725 (1997).
- Blocker, A. *et al.* *J. Cell Biol.* **137**, 113–129 (1997).
- Kuznetsov, S. A., Langford, G. M. & Weiss, D. *Nature* **356**, 722–725 (1992).
- Morris, R. L. & Hollenbeck, P. J. *J. Cell Biol.* **131**, 1315–1326 (1995).
- Rogers, S. L. & Gelfand, V. I. *Curr. Biol.* **8**, 161–164 (1998).
- Bi, G.-Q. *et al.* *J. Cell Biol.* **138**, 999–1008 (1997).
- Wu, X., Kocher, B., Wie, Q. & Hammer, J. A. III *Cell Motil. Cytoskel.* **40**, 286–303 (1998).
- Espresaco, E. M. *et al.* *Proc. Natl Acad. Sci. USA* **95**, 8636–8641 (1998).
- Lillie, S. H. & Brown, S. S. *Nature* **356**, 358–361 (1992).
- Alberts, B. *Cell* **92**, 291–294 (1998).

Reproductive biology

Deep-sea clams feel the heat

Cindy Lee Van Dover

Reproductive behaviours — the elaborate nesting rites of birds, cannibalistic courtships of praying mantis, lumbering love of elephants — are the centrepiece of television nature shows. The more remote the location, the more secretive the behaviour, the greater the hushed awe in the narrator's voice. By these measures, then, a narrator would be moved to virtual silence by video images of giant white clams spawning in the wild, more than 1,100 m below the surface of the sea. Such images have been documented by Fujiwara *et al.*^{1,2}, as they describe in *Deep-Sea Research*; even more remarkably, part of the project involved experiments at these depths.

The giant white clams, *Calyptogena soyoe*, of Sagami Bay, Japan (Fig. 1, overleaf), belong to the same genus as clams living in cracks between lobes of basalt lava at hydrothermal vents in the eastern Pacific. Like the vent clams, *C. soyoe* are host to endosymbiotic, chemoautotrophic, sulphide-oxidizing microorganisms, from which they derive their nutrition³. In Sagami Bay, the clams thrive in dense populations of hundreds to thousands in soft sediments where cold porewaters ('cold seeps'), rich in sulphide, migrate to the surface. The Sagami clams were the subject of a video and sensor observatory that gathered images and envi-

ronmental data over an 18-month period. During this time, spawning clams of both sexes were recorded 11 times, invariably associated with brief (less than two-hour) and hydrographically unexplained rises in temperature of about 0.1–0.2 °C in the overlying seawater, but uncorrelated with lunar or seasonal cycles.

What makes the new observations^{1,2} especially compelling is the experimental test of the hypothesis that the clams respond reproductively to slight increases in water temperature. Using the submersible *Shinkai 2000*, Fujiwara *et al.* placed a plastic dome equipped with a light bulb over an aggregation of clams and induced spawning within five minutes of turning on the light (which increased the temperature to 2.2 °C). At least five of the ten or so clams beneath the dome took part in this reproductive activity during more than one hour of observation. As is often the case in shallow-water bivalves, male clams spawned first. The neutrally buoyant gametes formed small clouds that drifted downstream, possibly stimulating females to spawn. No clams spawned in the control treatment.

It was not so long ago that the deep sea was viewed as the most unchanging environment on our planet, with a constant temperature and lack of food. With no external cues,



Figure 1 The giant white clams of Sagami Bay, which live and reproduce at depths of over 1,100 m.

deep-sea organisms were thought to have no periodicity in reproductive activity. In the past decade, however, through studies of seasonal pulses of photosynthetically derived detritus to the seabed and lagged but correlated periodicities in the reproductive condition of bottom-dwelling echinoderms⁴, the idea of the deep sea as a homogeneous environment has been abandoned. Even in this modern context, the work of Fujiwara and colleagues stands out as, to my knowledge, the first experimental demonstration of a direct link between a putative external cue and reproductive response in the deep sea.

The advantage of cued reproductive behaviour in species with broadcast spawning is patent: fertilization rates are increased where spawning is synchronized, and mixed pools of gametes and then larvae generated by mass spawning may find statistical refuge from predation and loss from the system by virtue of their larger numbers. The sensitivity of the Sagami clams to minute temperature changes may be an elegant reflection of the capacity for fine-tuning of animal to environment, and of the selective advantage

of reproductive synchrony in this species.

Based on histological studies, such synchrony has been inferred for several organisms inhabiting deep-sea hydrothermal vents as well⁵, begging the question of what the cues are in these ecosystems. Temperature might seem an unlikely possibility, given the already thermal nature of the system, but in fact predictable tidal signals are readily detected in time-series measurements at hydrothermal vents, with waxings and wanings of flow and thermal anomalies on diurnal and other lunar cycles⁶. Elaborations on the spawning experiments of Fujiwara *et al.* may allow resolution of some of the unanswered questions about reproductive behaviour and larval development in vent organisms. □

Cindy Lee Van Dover is in the Biology Department, Millington Hall, College of William & Mary, Williamsburg, Virginia 23187, USA.
e-mail: cindy_vandover@wm.edu

1. Fujiwara, Y., Tsukahara, J., Hashimoto, J. & Fujikura, K. *Deep-Sea Res.* **145**, 1881–1889 (1998).
2. <http://www.jamstec.go.jp/REMOTE/hatsushima/>
3. Endow, K. & Ohta, S. *Mar. Ecol. Prog. Ser.* **64**, 309–311 (1990).
4. Tyler, P. A. *Ocean. Mar. Biol. Annu. Rev.* **26**, 227–258 (1988).
5. Tyler, P. & Young, C. *J. Mar. Biol. Assoc. UK* (in the press).
6. Fujioka, K. *et al. Earth Planet. Sci. Lett.* **153**, 239–250 (1997).

Quantum control

Sculpting a wavepacket

Wolfgang P. Schleich

Thirty years ago, in a seminal article on the operational approach towards quantum mechanics, Willis E. Lamb¹ described the situation in quantum theory and measurement² by quoting Mark Twain: “Everybody talks about the weather but nobody does anything about it”. In the past few years, this situation has changed drastically — we now have precise instructions for two of the main stages of the measurement process in quantum mechanics, preparation and measurement of the quantum state (for an overview, see ref. 3). Another breakthrough is reported on page 233 of this issue by Weinacht *et al.*⁴, who describe an experiment in which they shape an atomic

electron’s wavefunction and drive it into a chosen state using tailored laser pulses (Fig. 1, overleaf). This work opens a new avenue in the field of coherent control and will have important applications in quantum computing and bond-selective chemistry.

Classical mechanics governs the motion of macroscopic objects; for example, the state and future motion of a tennis ball is completely determined by its position and momentum at a given time. In contrast, quantum mechanics is needed to describe the motion of microscopic objects, such as an electron in an atom. Heisenberg’s uncertainty principle forbids assigning both a well-defined position and momentum to the

electron. So, to determine completely the state of this quantum mechanical system, we need a whole distribution of position and momentum values for the electron (that is, the Wigner function).

Another way of expressing the state of a quantum mechanical system is the wavefunction, introduced in 1926 by Erwin Schrödinger. This quantity is a function of either position or momentum, and can take on complex values, having both real and imaginary parts. In the early days of quantum mechanics the physical interpretation of the wavefunction was subject to heated debate, and Schrödinger originally argued that only the real part of this quantity had physical significance. However, Max Born pointed out that only the absolute value squared of the wavefunction is directly experimentally accessible and corresponds to the probability of finding the position or momentum of a particle. (These historical discussions are summarized in ref. 2.) Since we can always represent a complex number by its amplitude and phase, this implies that only the amplitude can be measured directly. The phase is also needed to uniquely describe the quantum state, but how does one find this phase experimentally?

The question of ‘phase retrieval’ has a long experimental and theoretical history in classical optics, and in the field of quantum theory Wolfgang Pauli addressed it for the first time in 1933. Only recently, however, have techniques³ to reconstruct the complete wavefunction of a quantum mechanical system been implemented experimentally — in systems such as an ion, a diatomic molecule, an atomic beam or an electromagnetic field.

In a paper published last year, Weinacht *et al.*⁵, reconstructed the quantum state of a bound electron in a caesium atom. They used a variant of quantum state holography^{6,7} to measure the quantum state of the electron. Like classical holography, this technique relies on the interference between an object and a reference wave. In the case of a highly excited atom (a Rydberg atom), the object wave is the quantum state of the electron and the reference wave is a known state. In this experiment, which was a forerunner to the latest work, they illuminated the atom with two laser pulses derived from a common source (see lower section of Fig. 1). One pulse passed through a pulse shaper (box with dials), which allowed them to manipulate the amplitude, form and frequency of this pulse. Due to the different path lengths, there was a variable time delay between the two pulses. The first pulse (yellow) created an object wavefunction and the second one (red) served as a reference. Weinacht *et al.*⁵ measured the population of the individual energy eigenstates by applying a varying electric field across the condenser plates (shown in brown in Fig. 1). References 6 and 7 provide algorithms, which allow the reconstruc-

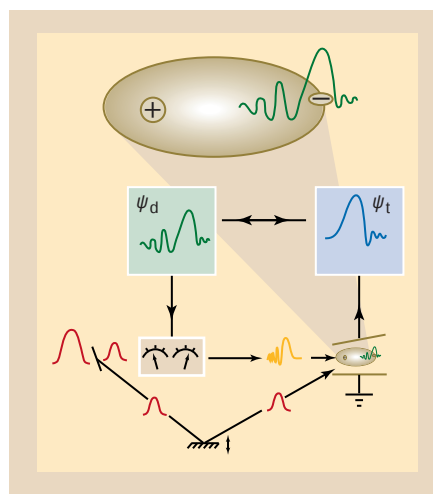


Figure 1 Sculpting an electronic wavefunction in an atom as described by Weinacht *et al.*⁴. The goal is to prepare the electron in a caesium atom (detail; top) in a desired wavefunction ψ_d (green). For this purpose they use two laser pulses (red and yellow) to first prepare a trial wavefunction ψ_i (blue) then reconstruct it and compare it to the desired state ψ_d . When the two differ they adjust the parameters of the preparation device (box with dials) and prepare a new trial state. The iteration of this cycle of preparation, readout and comparison leads to the desired wavefunction.

tion of the object wavefunction in amplitude and phase from the measured data.

However, before we can make a measurement of a quantum state we have to prepare that state. In the case of the caesium atom this job was done by the first object pulse. Can we prepare any desired state, that is, a target state? This is an issue that arises in many disciplines — ranging from physical chemistry, through atomic physics, to quantum optics — and many ways of tackling it have been suggested³. In the technique known as coherent control⁸, one tries to find the appropriate sequence of laser pulses needed to break a specific chemical bond, and in quantum state engineering one attempts to prepare quantum states of the radiation field or vibrational modes of an ion in a trap. Both approaches are central to recent proposals for quantum computers⁹.

How can we prepare a prescribed wavepacket of an electron orbiting the nucleus? The tools are short laser pulses. We therefore have to find the right shape and sequence of laser pulses to steer the quantum state to the target state. Usually the Schrödinger equation is used to calculate the quantum state that results after applying a sequence of laser pulses. However, in the case of coherent control, we concentrate on the inverse problem: we start from the target state and a rather simple initial state, and try to find the laser pulses that would lead to the desired state.

In their latest work, however, Weinacht *et al.*⁴ pursue a different approach and combine

quantum state preparation and measurement with feedback. The first laser pulse (yellow in Fig. 1) prepares a trial wavefunction, ψ_i (blue), and with the help of the second reference pulse (red) they reconstruct the trial state. The authors compare the reconstructed trial wavefunction ψ_i to the desired wavefunction ψ_d (green). When the two wavefunctions do not agree they readjust the dials on the pulse shaper so as to create a more appropriate laser pulse. This process is repeated until the desired wavefunction ψ_d emerges — usually within two cycles of the feedback loop.

The sculpting of wavepackets described by Weinacht *et al.*⁴, and the coherent control of electrons in a solid-state quantum well¹⁰ reported last year, are two impressive examples showing that the field of quantum control has entered a new era. It has broadened out beyond chemical reactivity and now allows encoding and decoding information on wavepackets. Moreover, the Weinacht procedure shows that it is possible to address

and control individual quantum units. This feature is crucial for the implementation of various suggestions for quantum computation and coherent information processing and transfer. The long-held dream of coherent control and quantum state engineering has become reality. □

Wolfgang P. Schleich is in the Abteilung für Quantenphysik, Universität Ulm, Albert Einstein Allee 11, D-89069 Ulm, Germany.
e-mail: schleich@physik.uni-ulm.de

1. Lamb, W. E. *Physics Today* **22**, 23–28 (1969).
2. Wheeler, J. A. & Zurek, W. H. (eds) *Quantum Theory and Measurement* (Princeton Univ. Press, 1983).
3. Schleich, W. P. & Raymer, M. J. *Mod. Opt.* **44**, 2021–2641 (1997).
4. Weinacht, T. C., Ahn, J. & Bucksbaum, P. H. *Nature* **397**, 233–235 (1999).
5. Weinacht, T. C., Ahn, J. & Bucksbaum, P. H. *Phys. Rev. Lett.* **80**, 5508–5511 (1998).
6. Leichte, C., Schleich, W. P., Averbukh, I. S. & Shapiro, M. *Phys. Rev. Lett.* **80**, 1418–1421 (1998).
7. Xin, C. & Yeazell, J. A. *Phys. Rev. A* **56**, 2316–2320 (1997).
8. Warren, W. S., Rabitz, H. & Dahleh, M. *Science* **259**, 1581–1589 (1993).
9. Braunstein, S. *Fortschr. der Phys.* **46**, 327–589 (1998).
10. Bonadeo, N. *et al. Science* **282**, 1473–1476 (1998).

Stress responses

Translational control perks up

Robert H. Silverman and Bryan R. G. Williams

The ability to respond to a variety of threats — from nutrient deprivation and viruses to chemical insults — is an essential property of cells. But how did primitive organisms survive these challenges, enabling them to evolve into complex animals? Stress responses are often linked to protein synthesis and folding which, in the budding yeast *Saccharomyces cerevisiae*, are controlled by two different kinds of gene, GCN2 and Ire1. On page 271 of this issue, Harding *et al.*¹ describe a newly discovered mammalian gene with features from both of these yeast genes.

In yeast, amino-acid starvation is sensed by the GCN2 protein, which phosphorylates a conserved serine residue on the α -subunit of the eukaryotic initiation factor-2 (eIF2 α)². Phosphorylation of eIF2 α suppresses translation from short upstream open reading frames within the messenger RNA for GCN4. The result is increased translation of GCN4, which activates transcription of the enzymes that synthesize amino acids.

Unfolded proteins in the yeast endoplasmic reticulum (ER) are sensed by Ire1, which is both a protein kinase and an endoribonuclease — an enzyme that cuts RNA molecules at internal sites^{3,4}. When unfolded proteins accumulate in the ER, Ire1 probably self-associates (oligomerizes) in the ER membrane, phosphorylates itself, then cuts an intron out of the precursor mRNA for a transcription factor, HAC1. After splicing and translation, the HAC1 protein induces

the transcription of several genes for proteins in the ER that mediate protein folding and modification.

GCN2 and Ire1 both need to be able to recognize RNA and phosphorylate other proteins. Compelling evidence from sequence analysis, backed up by experiment, indicates that during evolution the functional segments of the corresponding genes have been shuffled, allowing animals to cope with many different types of stress. In other words, GCN2 and Ire1 have evolved into a diverse family of mammalian stress-response proteins (Fig. 1).

Harding *et al.*¹ and Shi *et al.*⁵ have now cloned the gene for the newest player among these stress-response proteins, PERK (RNA-dependent protein kinase (PKR)-like ER kinase; also known as PEK, pancreatic eIF-2 α kinase). PERK combines functional properties of both GCN2 and Ire1, and it has sequence similarity to two mammalian homologues of these proteins, PKR and Ire1 β , respectively (Fig. 1). The domain related to Ire1 β is believed to detect unfolded proteins, whereas the PKR-related sequence encodes a protein kinase that can regulate protein synthesis. The amino-terminal, Ire1-related portion probably resides within the lumen of the ER, with the kinase domain in the cytoplasm. When cells are subjected to stress — such as unfolded proteins in the ER — eIF2 α is phosphorylated leading to a suppression of global protein synthesis, preserving energy and nutrients⁶. Harding *et al.* suggest that PERK, which is activated

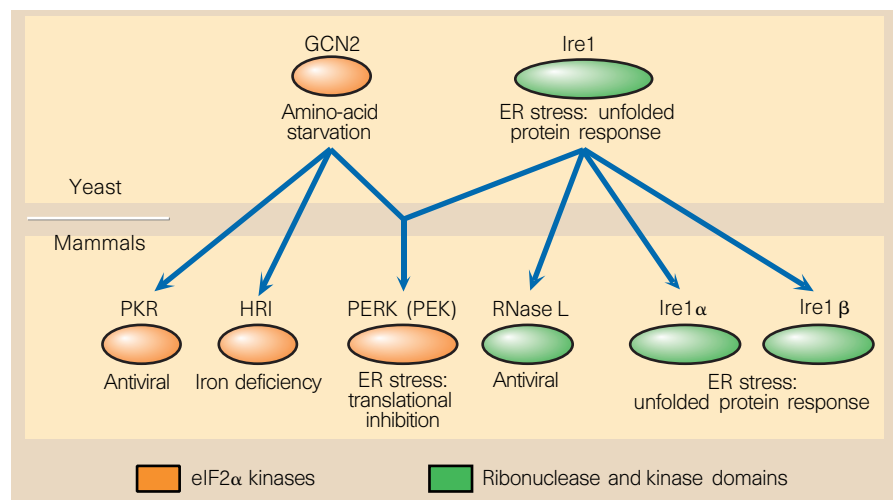


Figure 1 Relationship of yeast and mammalian stress-response proteins. The eukaryotic initiation factor (eIF2 α) kinase family originated with the yeast kinase GCN2, and shares homology with the mammalian RNA-dependent protein kinase (PKR), haem-regulated inhibitor (HRI), and the PKR-like ER kinase (PERK) cloned by Harding *et al.*¹ and Shi *et al.*⁵. Yeast Ire1 — both a kinase and an endoribonuclease — has homology with PERK, RNase L and the mammalian Ire1 α and β . So PERK has features of both GCN2 and Ire1, and is a missing link in the pathways.

during ER stress, is the protein most likely to be responsible for these effects.

Phosphorylation of eIF2 α is a key point in the regulation of protein synthesis in mammalian cells. Two kinases known to phosphorylate eIF2 α are PKR and the haem-regulated inhibitor (HRI) of erythroid cells^{2,7}. With GCN2, these proteins form a small family of eIF2 α kinases that contain a conserved eIF2 α recognition motif, amino-terminal to the kinase subdomain V. They can all regulate the phosphorylation of eIF2 α in yeast^{2,7}, a feature that Shi *et al.*⁵ also attribute to the ER stress-response kinase, PERK. Some of the effects of PKR mutants are probably due to inhibition of PERK. For example, although the ER stress response is normal in fibroblasts that lack PKR⁸, a *trans*-dominant PKR mutant in wild-type cells interferes with phosphorylation of eIF2 α in response to ER stress⁶. Moreover, the transforming activity of PKR mutants⁸ could implicate PERK in maintaining normal cellular homeostasis. Other properties of PKR may also apply to PERK. For example, PKR is targeted for inhibition by some viral proteins, allowing viruses to escape the antiviral effects of interferon⁸. The vaccinia virus K3L gene product contains regions of homology to eIF2 α , allowing it to inhibit PKR by mimicking its substrate — this should also inhibit the activity of PERK.

PERK lacks the conserved ribonuclease domain that is essential in yeast Ire1 and in three mammalian proteins, Ire1 α ⁹, Ire1 β ¹⁰ and RNase L^{11,12}. Ire1 α and Ire1 β are probably sensors and upstream effectors of the unfolded protein response. Why are there two Ire1 proteins in mammalian cells yet only one in yeast? It has been suggested that Ire1 α and Ire1 β might cleave at separate 5' and 3' splice sites in the RNA substrate,

unlike the yeast Ire1 which excises an intron from HAC1 pre-mRNA by cleaving at both sites^{3,9}. RNase L does an entirely different job. Type I interferons induce a family of proteins known as the 2', 5'-oligo A synthetases. When stimulated by double-stranded viral RNA, these proteins produce short, 2', 5'-oligoadenylates from ATP, which bind to inactive, monomeric RNase L, causing it to dimerize into its catalytically active form. The active RNase L digests single-stranded RNA, and RNA decay in interferon-treated cells attenuates viral replication⁸.

Although activation of Ire1 and RNase L is driven by unfolded proteins and 2', 5'-oligo A, respectively, both proteins oligomerize during activation. A functional link between RNase L, PKR and Ire1 β is that they are involved in mediating apoptosis induced by a variety of stimuli^{8,10}. Perhaps PERK may also be involved in apoptosis. Curiously, RNase L also has protein-kinase homology although, unlike Ire1, it lacks residues important for kinase function. Moreover, the amino-terminal regulatory regions of RNase L and Ire1 are unrelated, with ankyrin repeats and P-loop motifs in the 2', 5' oligo-A-binding domain of RNase L¹¹, and intraluminal and transmembrane domains in Ire1. Also, although the Ire1 proteins are highly specific RNases, RNase L has a broader specificity for UU and UA⁸.

By connecting two protein families involving eIF2 α kinase and ribonuclease activities, PERK stands as a kind of missing link in the evolution of stress-response proteins. The physiological roles of the mammalian stress-response proteins, and possible cross talk among them provide an exciting area for future study. □

Robert H. Silverman and Bryan R. G. Williams are in the Department of Cancer Biology of the Lerner



100 YEARS AGO

The performances of the submarine vessel, *Gustave Zédé*, appear to have given much satisfaction to naval experts on the other side of the Channel, though our own engineering papers are by no means impressed by the experiments. We learn from the *Times* that the semi-official *Moniteur de la Flotte*, commenting upon the trials of the *Gustave Zédé*, says that at length, after twelve years of continued efforts, the problem has been solved. The *Gustave Zédé*, unassisted, has steamed from Toulon to the Salins d'Hyères and to Marseilles ... and has successfully discharged her missiles at the mark. On the surface she is almost invisible, and presents a target scarcely capable of being hit; below water her presence is revealed neither by the noise of her engine nor any movement of the surface. The objection raised against the submarine boat that she is blind loses force, since the *Gustave Zédé* makes momentary appearances on the surface to redirect her course, while she has a telescopic tube, with an arrangement of prisms and mirrors, utilising the principle of the *camera obscura*, which permits the surroundings to be surveyed, though imperfectly, in case of emergency. The *Gustave Zédé* has restricted range, owing to the great weight of the electric accumulators; but the new boats of the Naval class will have auxiliary steam for surface navigation.

From *Nature* 19 January 1899.

50 YEARS AGO

Under the title "The Value of the Individual", Mr. F. I. G. Rawlins, in Occasional Paper No. 5 of the British Social Hygiene Council, asks physicists to look beyond their immediate pre-occupations. Physical science, he argues, arrives at a point where it can go further; but this does not justify the assumption that there is nowhere further to go. Modern physical theory cannot (with Laplace) postulate a universe which is a self-maintaining system, about the origins or destiny of which it is superfluous to inquire. The step from physics to theology is not compulsory; but there is nothing to prevent it and a good deal to encourage it. Only when that step is taken can the universe be seen as an environment with a meaning, where human personality is able to realize itself. From *Nature* 22 January 1949.

Research Institute, The Cleveland Clinic
Foundation, 9500 Euclid Avenue, Cleveland, Ohio
44195, USA.

e-mails: silvrr@cesmtp.ccf.org

wiliab@cesmtp.ccf.org

1. Harding, H. P., Zhang, Y. & Ron, D. *Nature* **397**, 271–274 (1999).
2. Dever, T. E. et al. *Proc. Natl Acad. Sci. USA* **90**, 4616–4620 (1993).
3. Sidrauski, C. & Walter, P. *Cell* **90**, 1031–1039 (1997).
4. Sidrauski, C., Chapman, R. & Walter, P. *Trends Cell Biol.* **8**,

245–249 (1998).

5. Shi, Y. et al. *Mol. Cell. Biol.* **18**, 7499–7509 (1998).
6. Brostrom, C. O., Prostko, C. R., Kaufman, R. J. & Brostrom, M. A. *J. Biol. Chem.* **271**, 24995–25002 (1996).
7. Chong, K. L. et al. *EMBO J.* **11**, 1553–1562 (1992).
8. Stark, G. R., Kerr, I. M., Williams, B. R. G., Silverman, R. H. & Schreiber, R. D. *Annu. Rev. Biochem.* **67**, 227–264 (1998).
9. Tirasophon, W., Welihinda, A. A. & Kaufman, R. J. *Genes Dev.* **12**, 1812–1824 (1998).
10. Wang, X. Z. et al. *EMBO J.* **17**, 5708–5717 (1998).
11. Hassel, B. A., Zhou, A., Sotomayor, C., Maran, A. & Silverman, R. H. *EMBO J.* **12**, 3297–3304 (1993).
12. Bork, P. & Sander, C. *FEBS Lett.* **334**, 149–152 (1993).

Animal behaviour

Monkey business in the aquarium

Rory Howlett

Occasionally, long-lost works come to light to the delight of scholars and laypersons alike — a sketch by Rembrandt, a Shakespeare sonnet or an early recording by the Beatles. So it is with the publication in *Evolution and Cognition* of a paper based on a manuscript drafted in February 1979 by Konrad Lorenz. The paper (K. Lorenz, K. Okawa & K. Kotschal *Evol. Cogn.* **4**, 108–135; 1998) has been completed by translator Kurt Kotschal and Keiko Okawa, a former student of Lorenz who has also provided additional results.

Along with Karl von Frisch and Nikolaas Tinbergen, Lorenz was a co-recipient of the 1973 Nobel Prize in Physiology or Medicine for their discoveries concerning “organization and elicitation of individual and social behaviour patterns”. At the time of his death in 1989, Lorenz had plans to write a book about “the biology, notably ethology of perciform fish”, on the basis of long-term observations of coral-reef fish. Observations on the competitive interactions within reef fish species had featured large in Lorenz’s classic text *On Aggression* (Harcourt Brace, 1963).

Many marine fish, including those that inhabit coral reefs, have larvae that are essentially planktonic, living at the mercy of ocean currents. A crucial stage is when they come out of the planktonic phase and settle on the reef. This recruitment phase is characterized by profound changes in morphology, behaviour and colouring. Lorenz was most struck by the vivid coloration of reef fish, and proposed that these colour patterns act as signalling ‘posters’ in the acquisition and defence of a territory. Others argued that the patterns might be involved in species recognition and mate choice, or defence against predators, either by way of camouflage or as warning signals.

Perhaps more than any other ethologist of his time, Lorenz recognized that to distinguish between these competing hypotheses, and to understand fully the social development of reef fish, detailed behavioural observation as well as experiment was required. In 1967, he spent some time in Hawaii making further notes on the behaviour of reef fish,

but evidently realized that what was needed was a large marine aquarium where the fish could be observed at leisure, and in which interactions between the fish could be experimentally manipulated.

The opportunity came in the mid-1970s, when Lorenz used money from his Nobel prize to construct a large reef tank at his home in Altenberg, Austria. This was no typical living-room tropical aquarium, but a giant 4 × 4 × 2-metre observation tank containing 32,000 litres of circulating sea water. The tank was stocked with the young of the Indo-Pacific coral-reef fish *Zanclus cornutus*, shown in the picture here, known as the Moorish idol or the *kihikihi* in Hawaiian. The first batch of fish did not prosper, but a second attempt to stock the tank was successful and between April 1976 and July 1978 Lorenz spent at least 1,000 hours observing the fish.

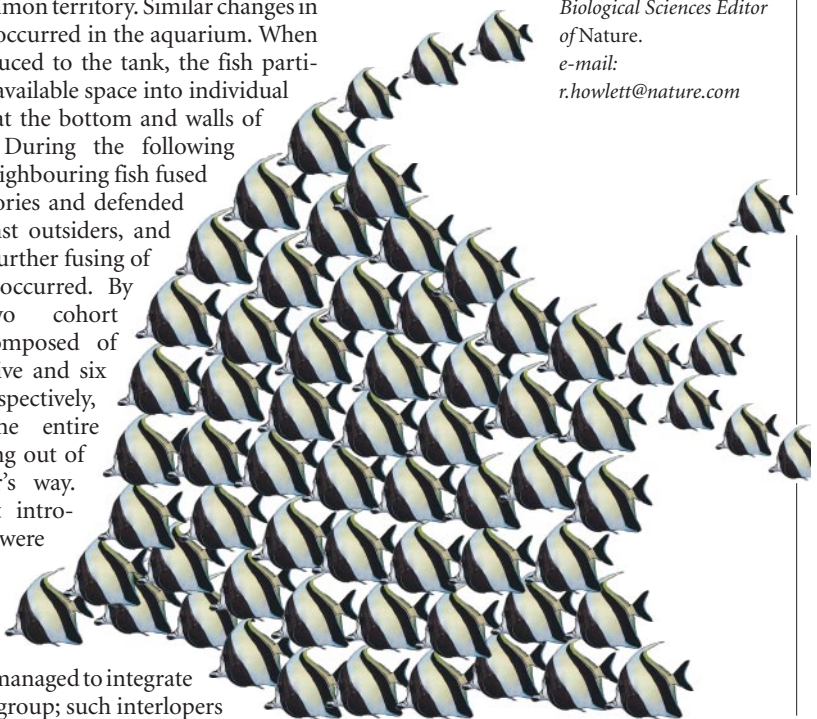
In the wild, newly recruited Moorish idols are territorial, defending their patch of reef. Later, individual territoriality breaks down and schools of more mature fish will often roam a common territory. Similar changes in behaviour occurred in the aquarium. When first introduced to the tank, the fish partitioned the available space into individual territories at the bottom and walls of the tank. During the following months, neighbouring fish fused their territories and defended them against outsiders, and with time further fusing of territories occurred. By 1978, two cohort groups, composed of fish aged five and six years respectively, roamed the entire tank keeping out of each other’s way. Subsequent introductions were attacked, often fatally, but sometimes they managed to integrate into a new group; such interlopers

occasionally sparked aggressive interactions between existing group members. Under natural conditions, such mechanisms might modulate group size to provide an optimal balance between warding off predators and competition for food, mates and so on.

The initial breakdown of individual territoriality fascinated Lorenz. During this period individuals seem to change the way in which they respond to ‘poster colour’ stimuli, depending on the ecological context and the motivational state of the fish. In the aquarium, at least, the character of the ‘dyadic’ relationships between pairs of fish provided evidence of individual recognition. Initially the interactions between fiercely territorial fish were aggressive. During the transition from strict territoriality to sociality, however, the dyadic relationships were characterized by a suite of ritualized behaviours apparently aimed at appeasement. These included parallel ‘side-by-side’ grazing along territorial borders, pseudospawning at the bottom of the tank, and eel-like swimming.

Complex appeasement behaviours are usually associated with highly social animals with well-developed cognitive abilities, such as primates. Lorenz’s findings demonstrated the potential for complex sociality in the Moorish idol, and similar behaviours are now known to occur in other coral fish, butterfly fish, for instance. In their commentary, Kotschal and Okawa liken the dynamic social structure of the Moorish idol to the ‘fission–fusion’ social behaviour of chimpanzees, which is characterized by dominance hierarchies and shifting coalitions of individuals. But whether the Moorish idols similarly ape primate social organization in the wild remains to be established. □

Rory Howlett is Deputy
Biological Sciences Editor
of *Nature*.
e-mail:
r.howlett@nature.com



Condensed-matter science

Ferroelectric ice

Steven T. Bramwell

Water molecules are dipolar, and while tumbling in the liquid state can be partly aligned by an electric field (Fig. 1). In ice, however, although the molecules are essentially static, there is no overall alignment (Fig. 2). A question that has long fascinated researchers is whether there is a form of normal ice in which the molecular dipoles have a net orientation towards one direction. Recent papers^{1–3} have revived an old debate about whether such ‘ferroelectric’ ice exists, and if so, what implications its existence would have for our understanding of condensed matter. The groups concerned report definite experimental evidence that at least partial ferroelectric alignment can be induced in normal ice, either by interaction with a substrate^{1,2}, or by doping with impurities³.

The debate dates back to the 1920s and the development of X-ray diffraction. Using this technique, it was established that the oxygen atoms of normal hexagonal ice (ice I_h) form a regular crystal structure in which each oxygen atom is coordinated with respect to four others in a tetrahedral arrangement (Fig. 2). X-rays are insensitive to hydrogen, but in 1933 Bernal and Fowler⁴ argued that each hydrogen must lie along the oxygen–oxygen line of contact, but displaced from the mid-point so as to form one shorter covalent bond and one longer hydrogen bond. To ensure that the structure consists entirely of water molecules, two hydrogens must be positioned closer to, and two others further away from, each oxygen. These ‘ice rules’ allow for many structures, differing in their relative orientations of the water molecules. In most ice structures the orientation of water molecules follows no regular pattern (Fig. 2), but Bernal and Fowler proposed that ice I_h has the simplest ordered arrangement — a form of ferroelectric ice in which the molecules are, on average, aligned.

They were undoubtedly guided in their choice of structure by the third law of thermodynamics, which states that the entropy of a perfect crystal is zero at a temperature of absolute zero (0 K). Entropy is proportional to the logarithm of the number of arrangements or motions available to the system of molecules, and so zero entropy means one arrangement ($\log 1 = 0$). It was suspected at the time (and still is) that the third law arose from a much deeper truth: that for any substance one definite crystal structure is the most stable, and is adopted at 0 K provided that equilibrium is attained. So to postulate a unique structure for ice seemed very reasonable.

It did not, however, prove to be correct. In the same year that Bernal and Fowler’s paper



Figure 1 Polarized water. In a simple experiment, a stream of tap water is deviated by the static electricity of a plastic pen (which has previously been charged by rubbing on wool) because the water molecules are partially aligned in the electric field.

appeared, Giaque and Ashley⁵ definitively demonstrated that ice still had entropy at 0 K. They did this by carefully comparing the ‘spectroscopic’ entropy for gaseous water (calculated from spectroscopic data using a theoretical expression), with the ‘third law’ entropy (estimated from calorimetric measurements by summing the entropy from 0 K upwards, and assuming the third law). The two estimates should be identical, but the third law entropy was $0.82 \pm 0.05 \text{ cal K}^{-1} \text{ mol}^{-1}$ (about 2% of the total) too small. Giaque and Ashley suggested that rotation of the water molecules might account for the discrepancy, but Pauling “consistently objected to” this explanation⁶. In a landmark paper in 1935 (ref. 7) he argued that ice could adopt any one of the huge number of molecular arrangements compatible with the Bernal–Fowler ice rules. He calculated this number and found it to be $(1.5)^N$, where N is the number of molecules — which, for a 1 g crystal of ice, is a number with 6×10^{21} digits! This translates to a residual entropy of $0.81 \text{ cal K}^{-1} \text{ mol}^{-1}$, exactly the value found experimentally (within error). Pauling’s explanation was immediately accepted⁶; but it was not until the advent of neutron diffraction, which can easily locate hydrogen in the form of its isotope deuterium, that the structure was directly confirmed^{8,9}.

It is now accepted that below about 160 K the water molecules of hexagonal ice settle into a random arrangement (Fig. 2), which is always preferred to an aligned one because there are so many more random structures to choose from. But in reality would an aligned structure be more stable? It may be that once ice has adopted a random form at high temperature, relaxation into the preferred state at low temperature is immeasurably slow. An

early report¹⁰ found a maximum in the electrical response of slightly impure ice at 100 K, suggestive of a transition to a ferroelectric state, but this was later ascribed to slow relaxational effects¹¹.

The latest work^{1,2} provides experimental evidence of some net ferroelectric alignment in films of ice grown on ultra-clean platinum surfaces. Ice films deposited at temperatures below about 120 K are amorphous, whereas those deposited at higher temperatures are crystalline, with either the hexagonal structure of the bulk, or a closely related cubic modification to which the Bernal–Fowler ice rules and Pauling’s arguments also apply (see caption to Fig. 2). Iedema *et al.*¹ use potential difference measurements to demonstrate a net polarization in amorphous or cubic ice films of 10^3 – 10^5 monolayers, deposited between 40 and 150 K. They argue that interaction with the substrate during growth aligns a surface layer of molecules, which then influence the orientation of water molecules in the rest of the sample. Only a small proportion (0.2%) of the molecules are actually aligned, but this is sufficient to give the sample a sizeable polarization, and may in fact be relevant to the agglomeration of ice particles in interstellar space¹.

Su *et al.*² show that surface interactions can indeed make ice ferroelectric: ultra-thin

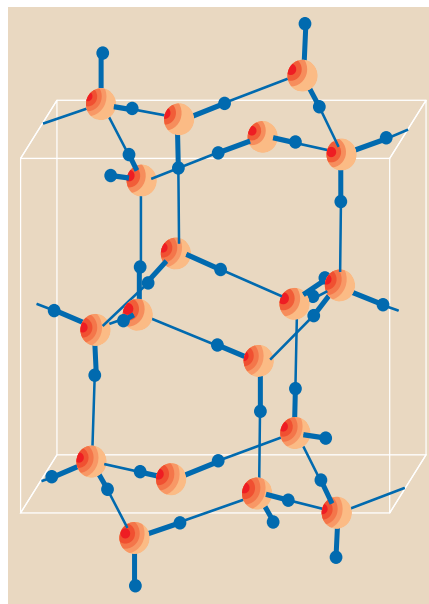


Figure 2 The structure of normal hexagonal ice I_h , showing the arrangement of oxygen atoms and one of the many possible random arrangements of hydrogen atoms (from ref. 9). The cubic modification of ice, I_c , has the same tetrahedral coordination of oxygen atoms, but the six-membered oxygen rings shown in the figure are repeated in a three-layer, rather than a two-layer, stacking sequence. Other high pressure phases of ice have different oxygen lattices and are called $II \dots X$. Three groups^{1–3} have demonstrated the existence of ferroelectric ice in which some of the dipolar water molecules are thought to be aligned.

hexagonal ice films (1–10 monolayers thick) have a net polarization that decays with distance from the substrate. They use 'sum frequency generation' to measure polarization in the sample — an ingenious method of overlapping a tunable-frequency infrared laser beam with a fixed-frequency visible laser beam. A vibrational spectrum at the summed frequency results only if the sample medium lacks inversion symmetry, a characteristic of aligned (ferroelectric) but not of random ice. The third approach to ferroelectric ice³ is to dope normal hexagonal ice with a catalytic quantity of hydroxide ions, a process that enhances the alignment (perhaps by enabling relaxation), without significantly disrupting the oxygen structure. Below 72 K this substance, known as ice XI, is suspected to be ferroelectric. Jackson *et al.*³ show that the neutron diffraction of ice XI is consistent with a ferroelectric arrangement, although the exact structure and degree of alignment are not unambiguously determined.

The experimental realization of ferroelectric ice would allow thermodynamic

measurements to establish whether nature really does prefer order at low temperature. However, as Iedema *et al.*¹ write "Over the years there have been many UFI citations (underidentified ferroelectric ices) in the literature", and it is not yet clear whether one of these elusive objects — a fully hydrogen-ordered ice — has finally fallen to earth. □

Steven T. Bramwell is in the Chemistry Department, University College London, 20 Gordon Street, London WC1H 0AJ, UK.

e-mail: s.t.bramwell@ucl.ac.uk

1. Iedema, M. J. *et al.* *J. Phys. Chem. B* **102**, 9203–9214 (1998).
2. Su, X., Lianos, L., Ron Shen, Y. & Somorjai, G. A. *Phys. Rev. Lett.* **80**, 1533–1536 (1998).
3. Jackson, S. M., Nield, V. M., Whitworth, R. W., Oguro, M. & Wilson, C. C. *J. Phys. Chem. B* **101**, 6142–6145 (1997).
4. Bernal, J. D. & Fowler, R. H. *J. Chem. Phys.* **1**, 515–548 (1933).
5. Giaque, W. F. & Ashley, J. W. *Phys. Rev.* **43**, 81–82 (1933).
6. Giaque, W. F. & Stout, J. W. *J. Am. Chem. Soc.* **58**, 1144–1150 (1936).
7. Pauling, L. *J. Am. Chem. Soc.* **57**, 2680–2684 (1935).
8. Wollan, E. O., Davidson, W. L. & Shull, C. G. *Phys. Rev.* **75**, 1348–1352 (1949).
9. Li, J. C. *et al.* *Phil. Mag. B* **69**, 1173–1181 (1994).
10. Dengel, O., Eckener, V., Plitz, H. & Riehl, N. *Phys. Lett.* **9**, 291–292 (1964).
11. Johari, G. P. & Jones, S. J. *J. Chem. Phys.* **62**, 4213–4223 (1975).

Telomeres

Capping off the ends

Carolyn Price

How many times have you had to deal with a loose end? Whether it be a stray end of hair, a frayed end of rope or the exposed end of a chromosome, ends can pose a problem. Reporting in *Cell*, Steve Schultz and co-workers¹ describe how the ciliate *Oxytricha nova* deals with the ends of its chromosomal DNA (telomeres). They have solved the crystal structure of the telomere end-binding protein (TEBP), which forms a protective cap over the DNA terminus. Their structure illustrates how the TEBP is exquisitely tailored to solve the chromosome-end problem², and also provides insights into how single-stranded (ss) DNA-binding proteins can recognize a particular sequence.

What is the chromosome-end problem? First, because DNA polymerase cannot copy the 5' end of a linear DNA molecule, chromosome ends must be replicated by some other mechanism. In most eukaryotes this is done by the enzyme telomerase, which adds tandem repeats to the DNA terminus. A second problem is that exposed DNA ends are lethal to cells. Not only do they trigger a DNA-damage response and subsequent cell-cycle arrest but, at chromosome ends, they allow degradation and end-to-end fusion. To avoid such dire events, the chromosome ends exist as DNA–protein complexes³.

The *Oxytricha* TEBP was the first telomere protein to be isolated^{3,4}. It is unusually abundant because *Oxytricha* has around 5×10^7 minichromosomes, and TEBP is found at each of the 1×10^8 telomeres. *Oxytricha*

telomeres consist of exactly 36 nucleotides of G_4T_4 sequence. The first 16 nucleotides protrude to form a single-stranded 3' $G_4T_4G_4T_4$ overhang, whereas the remainder is double stranded. The TEBP specifically recognizes the 3' overhang and binds tightly to form a protective complex over the end of the telomeric DNA. The protein is well-suited for its function as a chromosome cap because it binds only to the 3' terminus of single-stranded G_4T_4 DNA, and forms a very stable DNA–protein complex, protecting the DNA from nuclease digestion⁵. During DNA replication, the TEBP is thought to dissociate from the telomere to allow access to the replication machinery. On rebinding it may regulate the length of the telomere by displacing telomerase⁶.

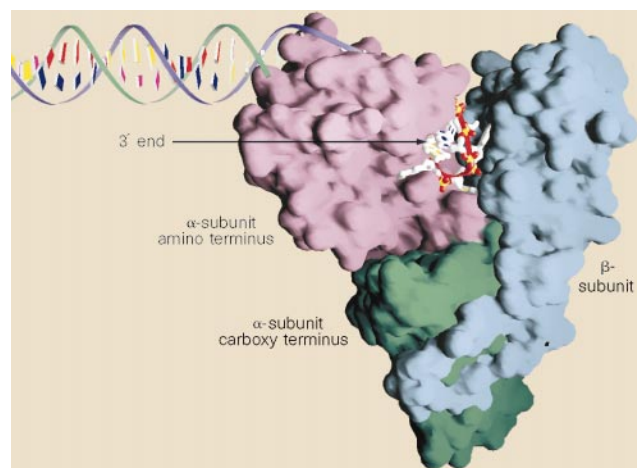


Figure 1 Crystal structure of the *Oxytricha* telomere end-binding protein (TEBP) bound to telomeric DNA. α - and β -subunits form a deep groove in which the single-stranded DNA sits. It seems that the DNA and protein must co-fold as part of the binding process, a new concept that could explain how ssDNA-binding proteins recognize a specific sequence.

subunit through an intricate set of hydrogen bonds and stacking contacts. Although Schultz and colleagues do not show the contacts between TEBP and the four most 5' Ts in the overhang, these residues probably also interact only with the α -subunit. The overall abundance of non-ionic interactions explains the unusual salt resistance of the native TEBP–DNA complex.

The intricacy of the protein–nucleic acid contacts and the deep burial of the ssDNA are striking features of the crystal structure. They suggest that, for the DNA to enter its binding site, the protein and ssDNA must co-fold as part of the binding process. If the protein surfaces in the groove between α and β were folded into a preformed structure, there seems to be no way that the ssDNA could enter its binding site. The concept that both the protein and DNA fold during binding is new, and may be an important aspect of sequence-specific recognition when proteins bind to single-stranded nucleic acids.

Because G-strand overhangs exist on the chromosomes of organisms ranging from slime moulds and yeast to humans and birds, a mechanism for protecting the overhang must be a standard requirement. Do other

organisms protect their chromosome ends with proteins that have a DNA-binding motif similar to that of the *Oxytricha* TEBP? Although none of the proteins identified by the various genome projects have sequence similarity to TEBP, some could nonetheless have a similar structure. This is because the oligonucleotide–oligosaccharide folds, the heart of the TEBP DNA-binding motif, show little sequence conservation to one another. In other words, proteins with completely different sequences might form equivalent structures at the telomere. □

Carolyn Price is in the Departments of Chemistry and Biochemistry, N146 Beadle Center, University of Nebraska, Lincoln, Nebraska 68588, USA.

e-mail: cmp@biocomp.unl.edu

1. Horvath, M. P., Schweiker, V. L., Bevilacqua, J. M., Ruggles, J. A. & Schultz, S. C. *Cell* **95**, 963–974 (1998).
2. Muniyappa, K. & Kironmai, K. M. *Crit. Rev. Biochem. Mol. Biol.* **33**, 297–336 (1998).
3. Gottschling, D. E. & Zakian, V. A. *Cell* **47**, 195–205 (1986).
4. Price, C. M. & Cech, T. R. *Genes Dev.* **1**, 783–793 (1987).
5. Price, C. M. in *Nucleic Acids and Molecular Biology* Vol. 9 (eds Eckstein, F. & Lilley, D. M. J.) 299–307 (Springer, Berlin, 1995).
6. Froelich-Ammon, S. J., Dickinson, B. A., Bevilacqua, J. M., Schultz, S. C. & Cech, T. R. *Genes Dev.* **12**, 1504–1514 (1998).
7. Gray, J. T., Celander, D. W., Price, C. M. & Cech, T. R. *Cell* **67**, 807–814 (1991).
8. Fang, G., Gray, J. T. & Cech, T. R. *Genes Dev.* **7**, 870–882 (1993).

pasteurianum hydrogenase the figure is $6,000\text{ s}^{-1}$ (see ref. 2). Extrapolation suggests that 1 mole of hydrogenase could produce enough hydrogen to fill the airship *Graf Zeppelin* in ten minutes, or the main liquid-hydrogen tank of the Space Shuttle in two hours (this fanciful calculation assumes a sufficient supply of reductant and protons, and disregards the time required to transfer hydrogen from solution to the gas phase).

Iron–sulphur proteins, which have clusters of sulphide and iron bound to cysteine residues of the protein, are widespread in most living cells. The redox potentials of many of them are low enough to produce hydrogen, but they do not do so. So what is different about the H-cluster which allows hydrogenase do to this?

The new protein structures reveal that, first, the H-cluster has two iron atoms with several diatomic ligands, believed to be cyanide and carbon monoxide. Such ligands are unusual in biology. But they have been observed in the [NiFe]-hydrogenases^{5,7}, and their presence in [Fe]-hydrogenases was foreshadowed by their infrared signature⁸. Second, there are free, or potentially free, ligand positions on the iron which could be occupied by a hydride or dihydrogen molecule. Third, although the catalytic site is deeply buried in the protein, it is accessible to hydrons.

The overall folding patterns of the two [Fe]-hydrogenases are similar, particularly around the active site. The *C. pasteurianum* enzyme has an additional domain containing two extra iron–sulphur clusters not seen in *D. desulfuricans*³, while the latter has a second subunit, without metal centres, forming a belt around the protein. Nicolet *et al.*⁴ propose that the second subunit is a feature of enzymes that are exported into the periplasmic space, and that a carboxy-terminal sequence which is cleaved from the large sub-

Bioinorganic chemistry

Hydrogenase sophistication

Richard Cammack

Hydrogenases, which catalyse the production and consumption of hydrogen gas, occupy a central place in the energy metabolism of anaerobic bacteria, and have probably done so since the earliest times of life on Earth¹. There are two types, one of which — the iron-only [Fe]-hydrogenases — have been especially difficult to study because of their sensitivity to oxygen².

Finally, however, structures of [Fe]-hydrogenases from *Clostridium pasteurianum* and *Desulfovibrio desulfuricans* have been determined, as reported by Peters *et al.*³ and Nicolet *et al.*⁴ in *Science* and *Structure*. In the layout of these proteins it is possible to recognize likely pathways for conduction of hydrogen, hydrons (H^+ ions) and electrons to the catalytic centre at the heart of the protein (Fig. 1). This centre of hydrogen production, known as the H-cluster, is revealed as an intricate structure of six iron atoms with a unique arrangement of nonprotein ligands.

Hydrogen is produced by many microbes under anaerobic conditions, and then consumed by others as a fuel^{1,2}. The two types of hydrogenase known are the nickel–iron ([NiFe]) hydrogenases, the first structure of which was solved in 1995 (ref. 5) and which are used mostly for hydrogen consumption; and the [Fe]-hydrogenases, which are most frequently used for hydrogen production. It turns out that the overall structures of the

[Fe]-hydrogenases are quite different to those of their [Ni–Fe] cousins, but some features are very similar and probably represent convergent evolution.

The [Fe]-hydrogenases are highly evolved catalysts. Under optimum conditions, each molecule of the *D. desulfuricans* enzyme can produce 9,000 molecules of hydrogen per second at 30 °C (ref. 6); for *C.*

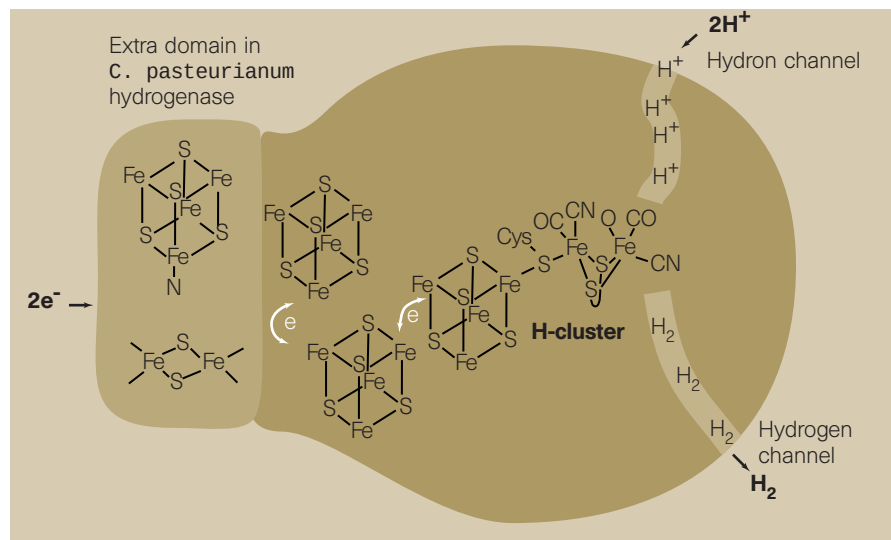


Figure 1 General features of the [Fe]-hydrogenases from *Clostridium pasteurianum* and *Desulfovibrio desulfuricans* (not to scale), as determined by Peters *et al.*³ and Nicolet *et al.*⁴. The enzymes catalyse the production of hydrogen in the reaction $2\text{H}^+ + 2e^- \rightleftharpoons \text{H}_2$.

unit may be a signal for cellular localization.

Other apparent differences between the enzymes lie in the two-iron part of the H-cluster. Both have two sulphurs positioned between the iron atoms. But in the 1.8-Å structure of *C. pasteurianum* hydrogenase, these are modelled as sulphides, with some additional electron density, and there is an extra bridging carbon monoxide; in the 1.6-Å structure of the *D. desulfuricans* enzyme, they are modelled as part of an organic ligand, propane-1,3-dithiolate.

A prominent feature of the two structures is the chain of iron-sulphur clusters leading to the H-cluster from the surface, where electron donors such as ferredoxins are expected to bind (Fig. 1). Iron-sulphur clusters, spaced at intervals of 1–1.5 nm, form an efficient electron-transfer chain, and their arrangement nicely illustrates the principle of modular design of electron-transfer proteins. Both hydrogenases have a domain containing two [4Fe-4S] clusters, which is structurally similar to those found in bacterial ferredoxins, and which could donate electrons to the H-cluster.

The *C. pasteurianum* hydrogenase additionally has two iron-sulphur clusters on a 'stalk' of protein extending from the globular domain. One is a [2Fe-2S] type, similar to those seen in plant ferredoxins. The other is a [4Fe-4S] cluster, with one iron being bound to histidine. A histidine-bound cluster is also seen in the [NiFe]-hydrogenase, but in *C. pasteurianum* the cluster is bound to the N epsilon of the imidazole ring of histidine, an unusual arrangement which points to convergent evolution.

Peters *et al.* and Nicolet *et al.* have defined the pathway for transfer of hydrons to the H-cluster in their respective structures, and it consists of amino-acid sidechains, capable of hydron exchange, along with protein-bound water molecules. A potential pathway for H₂ molecules has also been found; it seems that they do not simply diffuse through the protein, but are directed through specific channels. This was shown for the [NiFe]-hydrogenase by crystallography of the protein under high xenon pressure⁹ (the heavy xenon atoms were seen in pores leading through the protein to the active site, as predicted by computer modelling). In *D. desulfuricans* hydrogenase, Nicolet *et al.*⁴ see a narrow pore, which, with a little flexibility of the protein chain, would assist H₂ passage to the surface.

The H-cluster is one of a number of bioinorganic complexes, like the iron-molybdenum cofactor of nitrogenase, that owe their catalytic efficiency to a highly defined structure in a protein cavity. As with the H-cluster, several of these so-called 'superclusters'¹⁰ consist of a [4Fe-4S] cube linked covalently to another metal centre. These include sulphite reductase (cluster bridged through a cysteine ligand to a siro-

haem iron)¹¹; the P-clusters of nitrogenase (two fused [4Fe-4S] clusters)¹² and acetyl-CoA synthase (where there are two types of clusters, each bridged to a nickel centre)¹³. In these superclusters, the [4Fe-4S] group might serve as a capacitor to supply an electron during catalysis. But it remains to be seen how microbes are able to assemble such sophisticated pieces of coordination chemistry. □

Richard Cammack is in the Division of Life Sciences, King's College, London W8 7AH, UK.
e-mail: richard.cammack@kcl.ac.uk

1. Adams, M. W. W. & Stiefel, E. I. *Science* **282**, 1842–1843 (1998).

Neurobiology

A homeostatic switch

Richard Miles

It started life in French as *le contrôle du milieu interne*, and then picked up a Greek name — homeostasis. The idea, that in whole animals or single cells physiological parameters are regulated close to a constant value, is an old one. But homeostatic fixed points can change. A shift in the concentration of chloride during development changes the signalling between neurons that release the neurotransmitter GABA (γ-aminobutyric acid) and their target cells from positive to negative. And, reporting on page 251 of this issue, Rivera *et al.*¹ have pinned down a molecule that may be responsible for this switch.

On binding to its receptors, GABA opens channels that let through Cl⁻. In early life the concentration of Cl⁻ in nerve cells is set high, such that an efflux of Cl⁻ through these channels causes a depolarization (that is, the cell becomes more excitable; Fig. 1a, overleaf). But later, the fixed point for the intracellular concentration of Cl⁻ changes, and entry of Cl⁻ has the opposite effect, causing a hyperpolarization and reducing excitability (Fig. 1b). In this way, GABA is established as a major inhibitory neurotransmitter in the central nervous system.

A K⁺/Cl⁻ co-transporter seems to be responsible for this switch. This molecule couples the outward movement of Cl⁻ and potassium ions across the cell membrane, and has two known isoforms. One, KCC1, is a housekeeping protein expressed in most cells and involved in regulating cell volume. The other, KCC2, is specific to neurons², and is thought to be crucial for regulating the concentration of intracellular Cl⁻ (which is not distributed passively in these cells). Rivera and colleagues¹ now show that expression of KCC2 in rat hippocampal cells begins about one week after birth — just when GABA responses switch from depolarizing to hyperpolarizing. To consolidate the link they show that, in adult cells, they can change the hyperpolarizing action of GABA

- Adams, M. W. W. *Biochim. Biophys. Acta* **1020**, 115–145 (1990).
- Peters, J. W., Lanzilotta, W. N., Lemon, B. J. & Seefeldt, L. C. *Science* **282**, 1853–1858 (1998).
- Nicolet, Y., Piras, C., Legrand, P., Hatchikian, E. C. & Fontecilla-Camps, J. C. *Structure* **7**, 13–23 (1999).
- Volbeda, A. *et al.* *Nature* **373**, 580–587 (1995).
- Hatchikian, E. C., Forget, N., Fernandez, V. M., Williams, R. & Cammack, R. *Eur. J. Biochem.* **209**, 357–365 (1992).
- Happe, R. P., Roseboom, W., Pierik, A. J., Albracht, S. P. J. & Bagley, K. A. *Nature* **385**, 126 (1997).
- Pierik, A. J. *et al.* *Eur. J. Biochem.* **258**, 572–578 (1998).
- Montet, Y. *et al.* *Nature Struct. Biol.* **4**, 523–526 (1997).
- Pierik, A. J. *et al.* *Eur. J. Biochem.* **209**, 63–72 (1992).
- Crane, B. R., Siegel, L. M. & Getzoff, E. D. *Science* **270**, 59–67 (1995).
- Kisker, C., Schindelin, H. & Rees, D. C. *Annu. Rev. Biochem.* **66**, 233–267 (1997).
- Ragsdale, S. G. & Kumar, M. *Chem. Rev.* **96**, 2515–2540 (1996).

back to a depolarizing effect by injecting antisense RNA sequences that bind KCC2 messenger RNA and stop it being produced. The authors conclude that expression of KCC2 is enough to change the cell's fixed point for Cl⁻ from high to low.

Rivera *et al.* suggest that a similar switch from depolarizing to hyperpolarizing actions occurs even earlier in development of cells controlled by the other major inhibitory neurotransmitter, glycine. They found KCC2 at birth in neurons of spinal and hind-brain nuclei, indicating that the adult function — inhibition — is turned on sooner when mediated by glycine. But why do these inhibitory transmitters initially depolarize cells? In the developing cerebral cortex, GABA depolarizes immature neurons to the point where calcium enters through both voltage-gated and NMDA (N-methyl-D-aspartate) receptor-activated channels. The increased Ca²⁺ controls gene expression and neurotrophin signalling, so may contribute to the construction of axonal connections. In the hippocampus, GABA-mediated depolarization induces a synchrony of cell firing — a sort of precocious epilepsy³. The resulting large, rhythmic synaptic events and associated oscillations in intracellular Ca²⁺ may provide a natural trigger for plasticity of synaptic connections and, thus, for establishing and patterning the neural network.

Although some neurons express KCC2 early in development, and others later, many cells do not express this transporter at all. Is the level of intracellular Cl⁻ unregulated in these neurons? Certainly not — there are many other homeostatic mechanisms, including Cl⁻ channels, Cl⁻/bicarbonate exchangers, and transporters that couple movement of K⁺ and Cl⁻ to that of Na⁺. In Rohon-Beard cells⁴ and dorsal root ganglion neurons which, as Rivera *et al.*¹ show, do not express KCC2, the Na⁺/K⁺/Cl⁻ transporter sets Cl⁻ to a high level. This means that GABA responses are depolarizing, even in the

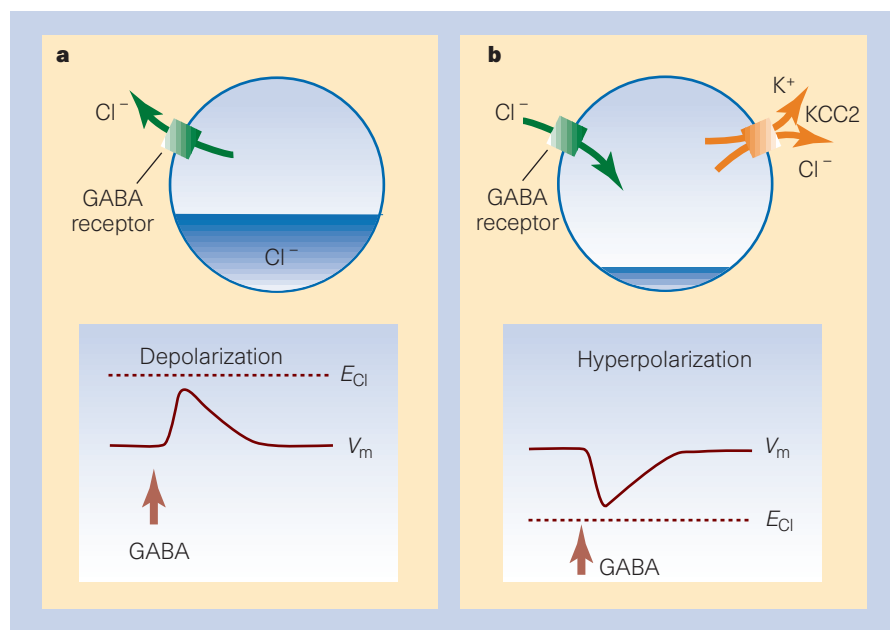


Figure 1 Chloride homeostasis and signalling by the inhibitory neurotransmitter GABA. **a**, Levels of intracellular Cl^- are high, and the equilibrium potential for Cl^- , E_{Cl} , is positive relative to the membrane potential, V_m . Opening of Cl^- channels by activation of the GABA_A receptor depolarizes the cell. **b**, Expression of the KCC2 transporter, studied by Rivera *et al.*¹, maintains a low level of Cl^- . E_{Cl} is negative relative to V_m and activation of the GABA_A receptor inhibits the cell. Immature hippocampal and spinal neurons start life as cells depolarized by GABA (**a**), and mature into cells in which GABA has a hyperpolarizing action (**b**). Dorsal root ganglion cells and some nerve terminals are depolarized by GABA into adult life (**a**).

adult. But, curiously, this excitatory action of GABA can inhibit synaptic signalling between neurons. The GABA acts on receptors located on synaptic terminals where Cl^- is set high. Depolarization⁵ prevents action potentials from being propagated into the nerve terminals, reduces the influx of Ca^{2+} and, in this way, inhibits transmitter release.

A depolarizing GABA action does not necessarily point to the absence of KCC2. Indeed, depolarization can be induced by quite another mechanism in adult hippocampal cells, where Cl^- is set low. The trick is that activation of a GABA receptor opens channels that are permeable to HCO_3^- as well as Cl^- (refs 6, 7). An influx of Cl^- tends to hyperpolarize, whereas efflux of HCO_3^- through the same channel depolarizes the cell. With prolonged stimulation, the Cl^- influx increases the concentration of intracellular Cl^- , reducing the hyperpolarizing action of GABA⁸. But because HCO_3^- is in a dynamic equilibrium with CO_2 , which can permeate the cell membrane, its depolarizing actions are maintained⁹. In this way, strong activation of GABA receptors can initiate biphasic responses — initially hyperpolarizing and then depolarizing — in adult neurons¹⁰.

Finally, how could regulation of a transporter molecule such as KCC2 contribute to neuronal plasticity? The KCC2 protein contains several consensus sites for phosphorylation, which are expected to regulate its activity². Might they, by modifying the homeostatic set point for the concentration

of intracellular Cl^- , underlie the circadian switch in GABA-mediated signalling¹¹ in the suprachiasmatic nucleus, from excitatory during the day to inhibitory at night? On a longer timescale, what would happen if KCC2 were downregulated, and how might this be triggered? Traumatic injury, for instance, causes a depolarizing shift in GABA responses with potentially pathological consequences¹²: the resulting influx of Ca^{2+} could kill cells, and remodelling of cortical circuits owing to the loss of inhibitory control could contribute to post-traumatic epilepsy. Certainly, homeostasis is not quite so static as it used to be. □

Richard Miles is in the Laboratoire de Neurobiologie Cellulaire, INSERM U261, Institut Pasteur, 25 rue du Dr Roux, Paris cedex 15, France.
e-mail: rmiles@pasteur.fr

1. Rivera, C. *et al.* *Nature* **397**, 251–255 (1999).
2. Payne, J. A., Stevenson, T. J. & Donaldson, L. F. *J. Biol. Chem.* **271**, 16245–16252 (1996).
3. Ben-Ari, Y., Cherubini, E., Coradetti, R. & Gaiarsa, J. L. *J. Physiol.* **416**, 303–325 (1989).
4. Rohrburgh, J. & Spitzer, N. C. *J. Neurosci.* **16**, 82–91 (1996).
5. Zhang, S. J. & Jackson, M. B. *Science* **259**, 531–534 (1993).
6. Kaila, K., Voipio, J., Paalasmaa, P., Pasternack, M. & Deisz, R. A. *J. Physiol.* **464**, 273–289 (1993).
7. Perkins, K. L. & Wong, R. K. S. *J. Neurophysiol.* **76**, 3886–3894 (1996).
8. Thomson, S. M. & Gahwiler, B. H. *J. Neurophysiol.* **61**, 501–511 (1989).
9. Staley, K. J., Soldo, B. L. & Proctor, W. R. *Science* **269**, 977–981 (1995).
10. Alger, B. E. & Nicoll, R. A. *Nature* **281**, 315–317 (1979).
11. Wagner, S., Castel, M., Gainer, H. & Yarom, Y. *Nature* **387**, 598–603 (1997).
12. van den Pol, A. N., Obrietan, K. & Chen, G. *J. Neurosci.* **16**, 4283–4292 (1996).

Daedalus

Cogito in vitro

The gloomy old orthodoxy about human brain cells was that, in the adult at least, they never divide or re-grow. Brain injury or deterioration was therefore irreversible. This belief now seems pessimistic. The human hippocampus, at least, can regenerate neurons; and mouse brain cells, encouraged by epidermal growth factor, have been successfully cultured *in vitro*. Last week Daedalus was exploiting these facts with his 'brain pan' — a single layer of brain cells cultured in a petri dish. The idea was for the cells to put out dendrites and axons, which would connect at random with each other. Electrodes lowered onto the cells could fire them, and eavesdrop on the resulting neural communication.

Daedalus now reckons that he has invented the ultimate neural-net computer. The 'strong' theory of neural-net and brain action, he says, is either brilliant or absurd, or possibly both. It asserts that a planless array of randomly wired neurons can learn anything. Neural connections which are reinforced become more easily triggered and more active; those which are neglected become less active. So an utterly planless network like his brain pan could be 'taught' to act as a computer. Even its wiring might grow to meet demand. Routes carrying heavy traffic could well grow more dendrites to carry it.

Daedalus is setting up a brain pan to test these claims. He is fitting it with a dense array of peripheral electrodes to simulate input (sensory) nerves, and another dense array of output (motor) connections. He will feed typical neural-net tasks into the input array, and assess the response on the output one. Correct responses will be reinforced by repetition; incorrect ones will be extinguished by neglect or counterexample. The first experiments will use mouse brain cells. Human ones might (or might not) be cleverer, but would raise ethical concerns.

The brain pan's education will start with simple signals for elementary evaluation, and will progress towards ever greater subtlety. Ultimately it will be given microphone and TV-camera inputs and hydraulic-actuator outputs, and encouraged to learn its way round the real world — like a mouse or human being. Success will raise deep philosophical problems about minds, meanings and consciousness. Failure will reinforce Daedalus's nagging suspicion that we know nothing whatever about how the brain works.

David Jones

Density-dependent warning coloration

Predators learn to avoid unpalatable prey more quickly when it is conspicuous than when it is cryptic¹. When they are rare, however, conspicuous individuals suffer a greater risk of discovery by predators than the cryptic majority, and their frequency is still too low for predators to learn to avoid them². We might therefore expect selection to favour unpalatable prey that is cryptic at low local densities and conspicuous at high local densities. I have found that such density-dependent warning coloration occurs in *Schistocerca emarginata* (= *lineata*) (Orthoptera: Acrididae) grasshoppers, providing insight into the biology of the desert locust, *Schistocerca gregaria*, and a new perspective on the evolution of warning coloration (aposematism).

Selection on warning coloration by visually hunting predators is predicted to result in monomorphic aposematic prey³. Despite this, polymorphisms for aposematism exist⁴. These polymorphic phenotypes represent variations on a similar aposematic theme, rather than a difference between crypsis and conspicuousness. Changes from cryptic to aposematic morphs have been thought of as evolutionary change, not polymorphism, but phenotypic plasticity in coloration can allow this transition to occur within a single generation.

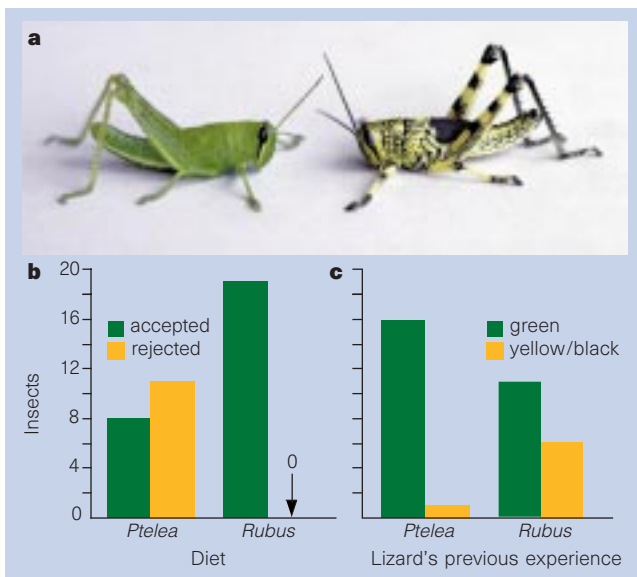
Some populations of juvenile *S. emarginata* in Texas feed mainly on *Ptelea trifoliata* (Rutaceae), despite the presence of other acceptable host plants⁵, causing nymphs to congregate there. *S. emarginata* nymphs from populations that feed on *Ptelea* exhibit a density-dependent colour polymorphism: typically, they are green at low rearing densities, and yellow and black at high rearing densities (Fig. 1a).

Juveniles from a *Ptelea*-feeding population were reared from hatching outdoors in screen cages as isolated individuals ($n=28$) or under crowded conditions in groups of ten ($n=31$). Increased rearing density significantly increased the melanized (black) proportion of the pronotum, wing, femur and abdomen of final-instar insects ($P<0.0001$, t -tests on arcsine-transformed data), as well as the frequency of yellow relative to green and yellow/green background coloration ($P<0.0001$, Monte Carlo contingency table). This polymorphism was independent of rearing diet⁵.

Yellow-and-black *S. emarginata* were unpalatable to insectivorous lizards (*Anolis carolinensis*) when reared on *Ptelea*, but not when reared on *Rubus trivialis* (Rosaceae) ($P<0.0001$, Fisher's exact test) (Fig. 1b). Unpalatability was mediated by gut content and was rapidly lost when insects were fed a non-toxic plant. All 14 *S. emarginata* that

Figure 1 Avoidance of *Schistocerca emarginata* by *Anolis carolinensis* lizards.

a, The green and yellow-and-black morphs of *S. emarginata*. **b**, *Ptelea*-conferred unpalatability. Adult *A. carolinensis* naive with respect to *S. emarginata* were controlled for motivational state and offered yellow-and-black *S. emarginata* nymphs reared on either *Ptelea* ($n=19$) or *Rubus* ($n=19$). Consumption without regurgitation was scored as acceptance. Attacking and releasing or regurgitation within 24 hours was scored as rejection. Each lizard was used once. **c**, Lizards discriminated against yellow-and-black in favour of green grasshoppers after experience with yellow-and-black *Ptelea*-reared grasshoppers. Lizards from the palatability assay were offered a simultaneous-choice test between a green *Chortophaga viridifasciata* nymph and a similar-sized yellow-and-black *S. emarginata* nymph. *C. viridifasciata* were used because it was not possible to rear green *S. emarginata* in sufficient numbers. Lizards eating both insects were omitted from the analysis.



were reared on *Ptelea* but fed Romaine lettuce for 24 hours were accepted by lizards ($P<0.0001$, Fisher's exact test, comparison with *Ptelea*-only diet treatment in Fig. 1b). Lizards previously fed unpalatable yellow-and-black *Ptelea*-reared *S. emarginata* discriminated against yellow-and-black grasshoppers in favour of green grasshoppers significantly more often than lizards previously fed palatable yellow-and-black *Rubus*-reared *S. emarginata* ($P<0.05$, Fisher's exact test; Fig. 1c). The yellow-and-black coloration of *S. emarginata* nymphs at high local densities can therefore function as a warning signal when the insects have been feeding on plants containing unpalatable chemicals.

The desert locust, *S. gregaria*, also exhibits a density-dependent nymphal colour polymorphism⁶, the biological role of which remains a mystery. My results indicate that the colour patterns of *S. gregaria* may be related to specific host plants. In non-outbreak populations, predation can be a strong mortality agent⁷. Nymphs eat plants containing compounds toxic to vertebrates⁸, and crowding on plants can facilitate colour change⁹. Density-dependent aposematism could reduce predation and population extinction rates in areas where insects feed on plants that make them unpalatable to predators, and could hasten the initial stages of population increase under favourable environmental conditions.

Phenotypic plasticity in coloration may have facilitated evolution of aposematism

in insects. Density-dependent changes from cryptic to conspicuous phenotypes occur in several insect lineages¹⁰. Among organisms with the appropriate genotype, density-dependent colour polymorphism could reduce the initial cost of conspicuousness once unpalatability (or some form of unprofitability) has been achieved. The evolution of this trait would be influenced by many factors, including, but not limited to, prey population dynamics and local predator behaviour. Examples of density-dependent aposematism may represent snapshots of this evolutionary process in action.

Gregory A. Sword

Department of Zoology and Brackenridge Field Laboratory, University of Texas, Austin, Texas 78712, USA
Present address: Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK
e-mail: greg.sword@zoology.oxford.ac.uk

- Gittleman, J. L. & Harvey, P. H. *Nature* **286**, 149–150 (1980).
- Mallet, J. & Singer, M. C. *Biol. J. Linn. Soc.* **32**, 337–350 (1987).
- Fisher, R. A. *The Genetical Theory of Natural Selection* 2nd edn (Dover, New York, 1958).
- Endler, J. A. *Phil. Trans. R. Soc. Lond. B* **319**, 505–523 (1988).
- Sword, G. A. *Host Plant Use and Density-Dependent Aposematism in Schistocerca (Orthoptera: Acrididae)*. Thesis, Univ. Texas (1998).
- Stower, W. J. *Anti-Locust Bull.* **32**, 1–75 (1959).
- Greathead, D. J. *J. Appl. Ecol.* **3**, 239–250 (1966).
- Bouaïchi, A., Simpson, S. J. & Roessingh, P. *Physiol. Entomol.* **21**, 247–256 (1996).
- Ghaout, S., Louveaux, A., Mainguet, A. M., Deschamps, M. & Rahal, Y. *J. Chem. Ecol.* **17**, 1499–1515 (1991).
- Fuzeau-Braesch, S. *Comp. Insect Physiol. Biochem. Pharmacol.* **9**, 549–589 (1985).

Fas-mediated cell death promoted by opioids

Opioids have been used as potent analgesics for centuries, but their abuse has deleterious physiological effects beyond addiction¹ — for instance, they can somehow alter the function of the immune system². Here we show that morphine induces the expression of the protein Fas (also known as CD95 or APO-1), a receptor on the cell surface that triggers the cell's suicide by apoptosis when it binds to its ligand, FasL. This induction of Fas expression by opioids appears to prime lymphocytes for elimination by apoptosis.

Several tissues and organs possess opioid receptors, but most of the information we have about their function comes from studies of neuronal cells. Lymphocytes have several different types of opioid receptor, making the immune system susceptible to perturbation by opioids².

Fas is a transmembrane protein that belongs to the family of receptors for tumour-necrosis factor and nerve-growth factor, and which helps to regulate the immune response^{3,4}; mice and humans with mutations in Fas or FasL both develop

lymphocyte-accumulation diseases⁴. As the Fas–FasL system is crucial for cellular homeostasis of the immune system, we reasoned that it might participate in the opioid-mediated alteration of the immune response.

The tightly regulated expression of Fas and FasL determines whether a lymphocyte survives or dies^{5,6}. We found that in a T-cell hybridoma (A1.1), in mouse splenocytes, and in freshly isolated human peripheral blood lymphocytes, morphine dramatically increases the expression of Fas (Fig. 1a) by activating the opioid receptors on these cells. This morphine-induced expression of Fas is blocked by naloxone, an antagonist of the opioid receptor.

Although morphine did not significantly affect the viability and proliferation of these cells, it primed them to undergo Fas-mediated apoptosis. Morphine-treated human peripheral blood lymphocytes underwent apoptosis when they were stimulated by L cells expressing Fas ligand (as indicated by the appearance of the hypodiploid peak in Fig. 1b). This morphine-primed, Fas-mediated apoptosis must be specific, because morphine-treated cells did not undergo apoptosis when co-cultured with L cells transfected with FasL in the antisense orientation. Results were similar when morphine-treated cells were

incubated with Fas-specific antibodies. In addition, morphine did not promote apoptosis in splenocytes from Fas-deficient mice, confirming that Fas is necessary for this process (data not shown).

It has been claimed that morphine can reduce the number of lymphocytes *in vivo*⁷, although in our *in vitro* experiments it was not sufficient to induce apoptosis. We confirmed this with splenocytes in mice *in vivo*, finding that morphine sulphate reduced the number of cells by 30% in 24 hours. We administered morphine together with either soluble Fas fusion protein (Fas-Ig) or normal mouse serum to test whether morphine caused loss of cells *in vivo* as a result of Fas–FasL interaction on these cells⁸: Fas-Ig prevented the morphine-induced loss of splenocytes, whereas normal serum did not (Fig. 1c). Morphine also induced Fas messenger RNA expression after 12 hours in the spleen, heart and lungs (Fig. 1d). As the morphine-induced loss of splenocytes *in vivo* is blocked by naloxone, this effect is presumably exerted through their opioid receptors, contributing to the immunosuppressive effects of morphine.

One way in which opioids perturb the immune system is by suppressing cytokine production, thereby inhibiting activation-induced proliferation. Our results indicate that they can also do this by directly inducing Fas expression and increasing apoptosis. Because Fas-mediated apoptosis requires FasL, which is expressed on a limited set of cells, only a small portion of cells undergo opioid-induced apoptosis under physiological conditions. But in pathological settings where the amount of FasL may increase, opioid-induced apoptosis could also increase.

Our discovery that opioids induce Fas expression not only bears upon the consequences of taking morphine, but should also help our understanding of the interaction between the nervous and immune systems.

Deling Yin, R. Allan Mufson*, Ruoxiang Wang, Yufang Shi

Department of Immunology,
Jerome H. Holland Laboratory for Biomedical
Sciences, American Red Cross,
15601 Crabbs Branch Way,
Rockville, Maryland 20855, USA
e-mail: shiy@usa.redcross.org

*Present address: Cancer Immunology Branch,
National Cancer Institute,
Bethesda, Maryland 20892, USA

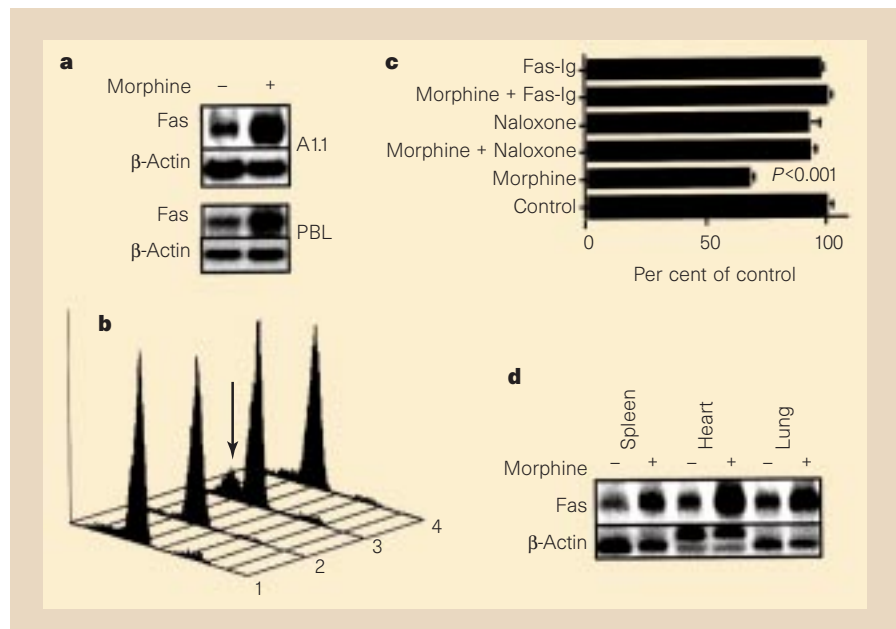


Figure 1 Morphine induces Fas expression and promotes FasL-mediated apoptosis. **a**, Morphine induces Fas expression. Northern-blot analysis of Fas expression in A1.1 T-cell hybridoma and human peripheral blood lymphocytes (PBL) treated with 3 μM morphine sulphate for 2 hours. **b**, Morphine primes human PBL to FasL-mediated apoptosis. PBL were treated with (peaks 1 and 3) or without (peaks 2 and 4) morphine for 2 hours and then co-cultured with mouse fibroblasts (L929) transfected with constitutively expressed FasL complementary DNA in the sense (peaks 3 and 4) or antisense (peaks 1 and 2) orientation. DNA content was analysed by flow cytometry after 12 hours. Arrow indicates the hypodiploid peak of apoptotic cells which was not evident in controls. **c**, The effects of *in vivo* administration of morphine sulphate on the cellularity of mouse spleens. BALB/c mice six weeks old were treated with morphine sulphate (50 mg per kg), Fas-Ig (150 μg per mouse) or naloxone (100 mg per kg). Total splenocytes were counted at 24 hours. **d**, Morphine injection induces Fas expression in different tissues. Fas expression was analysed by northern blotting 12 hours after morphine treatment.

- Roy, S. & Loh, H. H. *Neurochem. Res.* **21**, 1375–1386 (1996).
- Nestler, E. J. & Aghajanian, G. K. *Science* **278**, 58–63 (1997).
- Mariani, S. M., Matiba, B. & Krammer, P. H. *Behring Inst. Mitt.* **97**, 12–23 (1996).
- Nagata, S. & Golstein, P. *Science* **267**, 1449–1456 (1995).
- Shi, Y. F., Sahai, B. M. & Green, D. R. *Nature* **322**, 625–626 (1989).
- Brunner, T. et al. *Nature* **373**, 441–444 (1995).
- Stefano, G. B. et al. *Crit. Rev. Immunol.* **16**, 109–144 (1996).
- Suda, T., et al. *J. Immunol.* **154**, 3806–3813 (1995).

Does global warming make Triton blush?

Neptune's largest moon, Triton, is one of two satellites in the Solar System that are currently geologically active. At least two geyser-like plumes were observed by the Voyager 2 spacecraft in 1989, and dozens of streaky deposits hint at the existence of many more¹. Triton also exhibits complex seasonal changes in its 165-year journey about the Sun². Because Triton's atmosphere transports volatiles (primarily nitrogen and methane) during this seasonal cycle, its atmospheric pressure may fluctuate by up to an order of magnitude over decades³. Photometric measurements of its albedo and colour over half a century show that seasonal volatile transport has occurred⁴. There have also been indications that more extreme, short-lived changes, perhaps due to geological events, have occurred on Triton. An anomalously red spectrum was reported for Triton in 1977 (refs 5,6), and global warming has now been observed⁷.

We obtained four spectra of Triton between 0.35 and 0.95 μm using the 200-inch Hale telescope and the double spectrograph at Palomar Mountain Observatory on 23 October 1997. The spectral resolution was 10 Å between 0.35 and 0.55 μm , and 5 Å between 0.55 and 0.95 μm . We analysed the data according to standard procedures, and the 4-arcsecond slit was aligned to minimize the possibility of differential refraction. Our reduced spectrum for Triton is

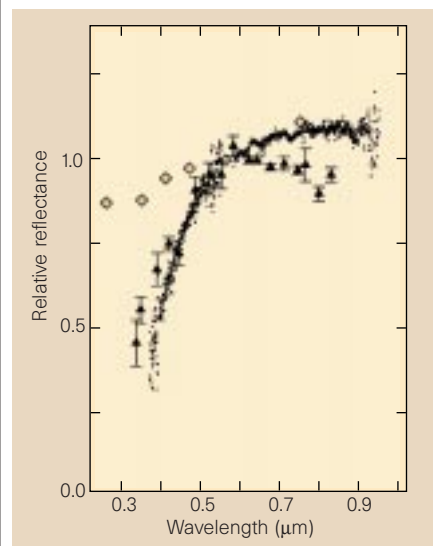


Figure 1 Spectral reflectivity of Triton. Our data were obtained in October 1997 at the Palomar Mountain Observatory (dots). The spectrum from the Voyager Imaging Science Subsystem and Photopolarimeter Subsystem (diamonds) is representative of Triton's colour in 1990 and just before, whereas the 1977 data^{5,6} (triangles) represent a previous observation of a reddening of Triton's spectrum.

compared with previous observations in Fig. 1: it is significantly different from that normally observed, but is similar to the anomalous 1977 spectrum^{5,6}. The upper limit for the timescale for our observed spectrum is seven years (the period between the last published spectrum in the range 0.35–0.55 μm (ref. 4) and October 1997).

A global temperature increase of nearly 2 K was observed on Triton between 1989 and November 1997 (ref. 7). Our observations indicate that this change may be due to deposition of material from a geological process on Triton. Both the spectral changes and the global warming may have been caused by a triggering event, such as massive venting, causing the deposition of dark, red material onto the surface of Triton. The spectrum of the streaks that appear to be plume deposits are indeed darker and redder than Triton as a whole⁸. The most volatile, fresh, blue nitrogen ice may initially sublimate to expose a lower layer of red photolysed methane ice⁹, perhaps causing an additional temperature rise. Of the mapped geological regions, our spectrum is most like that of a region that is probably rich in irradiated methane clathrate⁸. Our spectrum has a more pronounced methane absorption band at about 0.73 μm than a spectrum obtained in 1989 (ref. 10).

Whatever the triggering mechanism(s), Triton's surface will warm significantly as its visual and near-ultraviolet albedo decrease. Triton's Bond albedo ($A = pq$, where p is the geometric albedo and q is the phase integral) is among the highest of any body in the Solar System. Its energy balance is therefore particularly sensitive to small variations in its spectral albedo. At the time of the Voyager 2 encounter, A was 0.89 in the visual region of the spectrum, where most of the solar flux occurs¹. Because $T \approx (1 - A)^{1/4}$, where T is the temperature, a decrease of only about 3% in the Bond albedo would account for the observed temperature increase⁷. Increases in the absolute (geometric) albedo, which our observations were not designed to measure, may mitigate some of the decrease in the Bond albedo.

Models for global seasonal change on Triton that include only the effects of solar insolation could underestimate global temperature changes and the atmospheric loading of volatiles.

Bonnie J. Buratti, Michael D. Hicks, Ray L. Newburn Jr

Jet Propulsion Laboratory,
California Institute of Technology,
Palomar Observatory,
4800 Oak Grove Drive 183-501,
Pasadena, California 91109, USA
e-mail: bburatti@ocelot.jpl.nasa.gov

1. Smith, B. *et al.* *Science* **246**, 1422–1449 (1989).
2. Trafton, L. *Icarus* **58**, 312–324 (1984).
3. Brown, R. H., Johnson, T. V., Goguen, J. D., Schubert, G. & Ross, M. N. *Science* **251**, 1465–1467 (1991).

4. Buratti, B., Goguen, J. D., Gibson, J. & Mosher, J. *Icarus* **110**, 303–314 (1994).
5. Cruikshank, D. P., Stockton, D., Dyck, H. M., Becklin, E. E. & Macy, W. *Jr Icarus* **40**, 104–114 (1979).
6. Bell, J., Clark, R. N., McCord, T. B. & Cruikshank, D. P. *Bull. Am. Astron. Soc.* **11**, 570 (1979).
7. Elliot, J. *et al.* *Nature* **393**, 765–767 (1998).
8. McEwen, A. *J. Geophys. Res.* **17**, 1765–1768 (1990).
9. Thompson, W. R., Murray, B., Khare, B. N. & Sagan, C. *J. Geophys. Res.* **92**, 14933–14947 (1987).
10. Grundy, W. & Fink, U. *Icarus* **93**, 379–385 (1991).

Shikimate pathway in apicomplexan parasites

The discovery of plastids in apicomplexan parasites^{1–3} raised the possibility that these organelles might harbour plastid-specific metabolic activities that could be blocked by therapeutic agents. Ideally, these agents would inhibit the parasites without harming their vertebrate hosts. Roberts *et al.*⁴ have made the promising discovery that Apicomplexa are sensitive to the herbicide glyphosate (better known by its trade names of RoundUp, Zero or Tumbleweed), which is an inhibitor of the enzyme 5-enolpyruvyl shikimate 3-phosphate synthase. They suggested that production of aromatic amino acids by the pathway involving this enzyme, the shikimate pathway, might be an essential function of the apicomplexan plastid, but here we present evidence that this pathway actually operates in the cytosol of Apicomplexa.

The shikimate pathway occurs in prokaryotes, fungi and the plastids of plants and algae, but not in vertebrates⁵. In plants, the shikimate-pathway enzymes, like other nuclear-encoded plastid proteins, are post-translationally targeted to the plastid by means of an amino-terminal leader sequence. Apicomplexan proteins destined for the plastid also include substantial leader peptides, which are sufficient for the transport of reporter proteins across all four membranes of the plastid of the apicomplexan parasite *Toxoplasma*⁶.

In contrast to the plant enzymes, there is no evidence that apicomplexan shikimate enzymes are localized in the plastid, and neither do the *Toxoplasma* and *Plasmodium* chorismate synthase genes encode leader peptides (see Fig. 2 of ref. 4). This absence of a leader strongly suggests that the apicomplexan chorismate synthase proteins function in the cytosol and not the plastid. This would not be unprecedented, for the shikimate pathway operates in the cytosol of fungi⁷. Moreover, the fungal shikimate pathway is sensitive to glyphosate⁷.

Phylogenetic analysis of chorismate synthase (Fig. 1) is consistent with this idea, as the apicomplexan proteins are most closely related to fungal cytosolic proteins, whereas plastid-targeted proteins from plants group

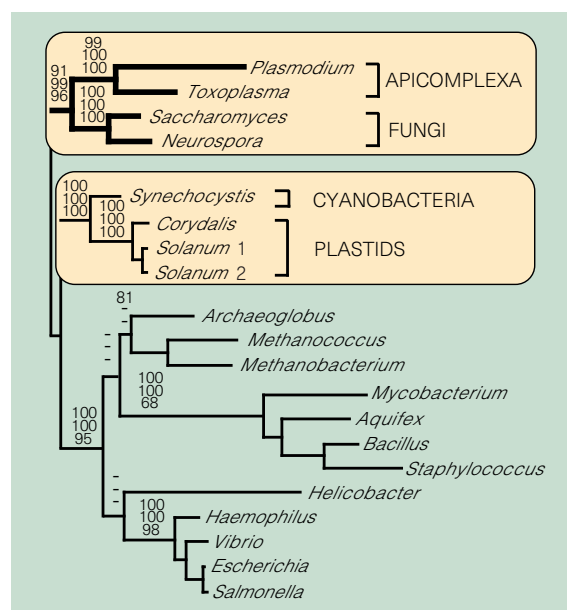


Figure 1 Phylogeny of chorismate synthase. The figure shows a Fitch-Margoliash tree, based on maximum-likelihood distances corrected by the Jones, Taylor and Thornton substitution matrix. Other methods (neighbour-joining, maximum parsimony, protein maximum likelihood, and quartet puzzling) all also support the grouping of fungi and Apicomplexa, and of cyanobacteria and plastids. Numbers at the nodes correspond to support for the major groups from (top to bottom): Fitch-Margoliash bootstrap, protein maximum-likelihood resampling estimated log-likelihood bootstrap, and per cent occurrence in quartet-puzzling steps. Dashes indicate support of less than 50 per cent.

with cyanobacteria, as expected. The grouping of the fungal/apicomplexan lineage with the cyanobacterial/plastid lineage, although curious, should not be mistaken for plastid origin because this tree is unrooted, the cytosolic genes do not branch specifically with plastids, and fungi lack plastids and indeed probably never had them.

The evolutionary origin of cytosolic shikimate-pathway enzymes in eukaryotes is unclear, as the distribution of the cytosolic pathway in eukaryotes is poorly understood and the sampling of sequences is limited to fungi and Apicomplexa. Nevertheless, in view of the phylogeny of chorismate synthase and the absence of leader peptides, there is no reason to believe that the shikimate pathway of Apicomplexa is either derived from or functions in the plastid.

Even if the shikimate pathway does take place in the cytosol in Apicomplexa, this does not detract from the importance of the sensitivity of these parasites to glyphosate. Vertebrate hosts of Apicomplexa lack this pathway, so it remains a specific chemo-therapeutic target. However, before glyphosate sensitivity can be exploited, we need to define the role, if any, of the plastid in the apicomplexan shikimate pathway.

Patrick J. Keeling*, **Jeffrey D. Palmer***, **Robert G. K. Donald†**, **David S. Roos‡**, **Ross F. Waller‡**, **Geoffrey I. McFadden‡**

*Department of Biology,
Indiana University,

Bloomington, Indiana 47405, USA

†Department of Biology,

University of Pennsylvania,
Philadelphia, Pennsylvania 19104, USA

‡Plant Cell Biology Research Centre,
Botany School, University of Melbourne,
Parkville, Victoria 3054, Australia

3. Köhler, S. *et al. Science* **275**, 1485–1489 (1997).

4. Roberts, F. *et al. Nature* **393**, 801–805 (1998).

5. Kishore, G. M. & Shah, D. M. *Annu. Rev. Biochem.* **57**, 627–663 (1988).

6. Waller, R. F. *et al. Proc. Natl Acad. Sci. USA* **95**, 12352–12357 (1998).

7. Boocock, M. R. & Coggins, J. R. *FEBS Lett.* **154**, 127–133 (1983).

Roberts et al. reply — Although our study¹ was stimulated by the discovery of the apicomplexan plastid, we did not claim to have provided proof that the shikimate pathway occurred in the plastid in apicomplexans. We believe that the site of this pathway remains an open question, despite the similar data emerging from our own² and Keeling *et al.*'s phylogenetic analyses of our chorismate synthase gene sequences.

The premise for the phylogenetic analyses was that if the shikimate pathway occurs in the plastid, then these gene sequences are likely to show evolutionary relatedness to those in green algae, like the genes present on apicomplexan plastid DNA. However, phylogenetic analysis of chorismate synthases in the apicomplexans does not support this hypothesis; rather, it points to a closer similarity with fungal enzymes. From this, Keeling *et al.* conclude that the shikimate pathway is cytosolic (as it is in fungi).

This situation might have been expected if apicomplexans had evolved directly from a common fungal lineage, but this is not generally believed to be the case. At present, the best evidence available aligns apicomplexans more closely with ciliates and plants, with fungi as distant relatives.

Caution is needed in drawing conclusions about the subcellular locations of proteins from phylogenetic analysis. Such studies are fraught with difficulty because of the frequency with which horizontal gene transfer occurs during evolution, especially when the data are from a relatively small number of organisms. We believe that phylogenetic analysis is an inadequate tool for

determining the location(s) of the shikimate pathway in apicomplexans: direct experimental evidence of the location of key proteins is required.

The two apicomplexan chorismate synthases do not have an obvious amino-terminal leader sequence like that found on plastid chorismate synthases of higher plants. However, apicomplexan protein sequences differ from all other reported chorismate synthases in that they have a number of large insertions. The function(s) of these insertions remain(s) unknown, but one of them might act as a targeting sequence, as proteins can be targeted to plastids even without an amino-terminal leader sequence³.

The location of the shikimate pathway in apicomplexan parasites remains an open question. A rigorous, practical approach should reveal that all, some or even none of the shikimate enzymes function within the plastid, or the pathway may turn out to operate both outside and inside the plastid in apicomplexans, as in *Euglena*⁴. Irrespective of where the pathway is located, however, it presents a prime target for antiparasitic agents. A cytosolic location would be more accessible to a target enzyme as fewer membranes would need to be crossed.

Craig W. Roberts*†‡, John Finnerty\$, Jennifer J. Johnson*, Fiona Roberts*†, Dennis E. Kyle||, Tino Krell¶, John R. Coggins¶, Graham H. Coombs¶, Wilbur K. Milhous||, Saul Tzipori#, David J. P. Ferguson☆, Debopam Chakrabarti|, Rima McLeod*

*Michael Reese Hospital and Medical Center,
Chicago, Illinois 60616, USA

and the University of Chicago,
Chicago, Illinois 60637, USA

‡Department of Immunology,
University of Strathclyde, Glasgow G4 0NR, UK

\$Department of Organismal Biology and Anatomy,
University of Chicago,
Chicago, Illinois 60637, USA

||Walter Reed Army Institute of Research,
Washington DC 20307-5100, USA

¶Institute of Biomedical and Life Sciences,
University of Glasgow, Glasgow G12 8QQ, UK

#Tufts University School of Veterinary Medicine,
Boston, Massachusetts 01536, USA

☆John Radcliffe Hospital, Oxford University,
Oxford OX3 9DU, UK

|University of Central Florida,
Orlando, Florida 32816-2360, USA

†Present addresses: University of Strathclyde,
Glasgow, UK (C.W.R.); University of Glasgow,
Glasgow, UK (F.R.)

1. Roberts, F. *et al. Nature* **393**, 801–805 (1998).

2. Roberts, C. W. *et al. Keystone Symposia on Molecular and Cellular Biology, Keystone, Colorado, April 2–8, 1998* (Keystone, Silverthorne, 1998).

3. Hugosson, M., Nurani, G., Glaser, E. & Franzen, L.-G. *Plant Mol. Biol.* **28**, 525–535 (1995).

4. Reinbothe, C., Ortel, B., Parthier, B. & Reinbothe, S. *Mol. Gen. Genet.* **245**, 616–622 (1994).

1. Wilson, R. J. *et al. J. Mol. Biol.* **261**, 155–172 (1996).

2. McFadden, G. I., Reith, M. E., Munholland, J. & Lang-Unnasch, N. *Nature* **381**, 482 (1996).

Betwixt hagiography and debunking

Louis Pasteur

by Patrice Debré

Johns Hopkins University Press: 1998. 550 pp.
\$39.95, £23.74

W.F. Bynum

During the past couple of decades, while historians have been fretting over Darwin's neuroses, Newton's mental instability and Einstein's sex life, Pasteur has also received his share of attention. Most commentary has perpetuated the traditional cult of the great man, an attitude already well established during Pasteur's lifetime and capped by the French declaration of 1995, the centenary of his death, as *l'année Pasteur*. But amid the celebrations of this most iconic of scientists, a discordant note was sounded by Gerald Geison's *The Private Science of Louis Pasteur* (Princeton University Press, 1995), which exploited Pasteur's laboratory notebooks to argue that the great man was more devious, and more ruthless, than his studied public persona revealed.

Patrice Debré's biography was originally published in 1994, just in time for the centennial festivities, and he has contented himself in this English edition with merely distancing himself from the interpretations that Geison ("the American historian") placed on the episodes he concentrated on. Debré also hints darkly that the reception of Geison's work in France was coloured by Franco-American priority disputes about the discovery of the AIDS virus.

In ignoring Geison, Debré, an immunologist who works on AIDS, has not lived up to the model provided by his subject, who loved nothing better than to refute his opponents point by point. This new edition would have given Debré the opportunity to reflect on Geison's work; instead we are offered merely a short bibliography of reviews stimulated by "*l'affaire Geison*", without even a citation of Geison's monograph. This was not Pasteur's way of confronting controversy.

For all that, Debré promised a biography that would lie between hagiography and debunking. He certainly achieves that for Pasteur the man: his portrait of this taciturn, secretive individual does not make his subject someone likely to rank high on anyone's list of historical figures with whom one would like to have dinner. Even Pasteur's devotion to family life was diluted by his obsession with work, around which all else revolved. He took his wife into his confidence, probably because Marie Laurent Pasteur acted as his principal secretary. His co-workers were as often as not set tasks the significance of which was known only to Pasteur himself.

By contrast, Pasteur the scientist

emerges more or less unsullied. Some of the issues that Geison highlighted, such as the influence of Auguste Laurent on Pasteur's early crystallography experiments, are barely mentioned. Others (the source of the method for making anthrax vaccine, or the secret attempts at rabies vaccination before Joseph Meister) are reinterpreted. Pasteur's critics, such as Michel Peter (dubbed by Debré as "the Man Consumed by Envy"), are subtly marginalized. Another opponent, Jules Guérin, is described as "prattling" when asking Pasteur for more detail about his theories of vaccines and immunity. Guérin actually challenged Pasteur to a duel.

At times, Pasteur's triumphs are exaggerated. Even while acknowledging how difficult the rabies vaccine was to evaluate, Debré writes that four American children, bitten by a supposedly rabid dog in New York City, "arrived in Paris in time to be saved". Constructing the story like this neutralizes the valid point of those who were dubious of Pasteur's rabies vaccine — that 'success' was impossible to evaluate. There is simply no way of knowing now, nor was there then, what would have happened to the children had they not journeyed to Paris to receive Pasteur's rabies vaccine.

Throughout, Debré is too content to see the world only through Pasteur's eyes. For example, his history of the spontaneous-generation controversy is reconstructed almost entirely through Pasteur's historical passages, to the neglect of recent scholarship by Farley, Geison, Secord and others.

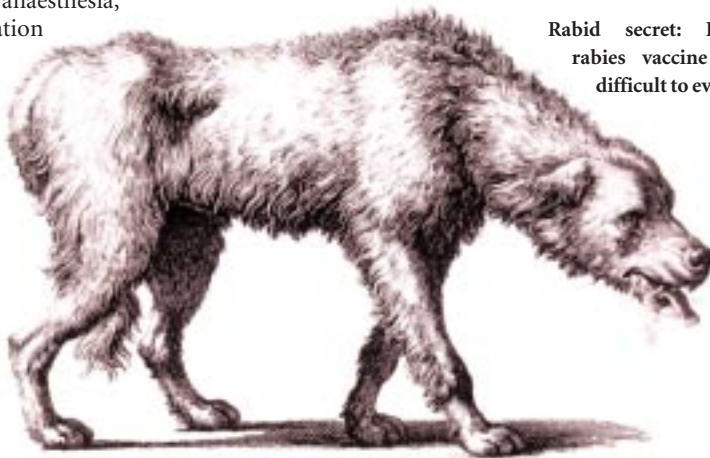
Scene-setting pages frequently betray careless research. Dates, such as Lamarck's coining of the term '*biologie*', the introduction of anaesthesia, the publication of Claude

Bernard's *Introduction to the Study of Experimental Medicine*, or the death of Jules Verne, are simply wrong. Ignaz Semmelweis's work is retold through the romantic biography by Louis-Ferdinand Céline, now more than half a century out of date. Joseph Lister's work is crudely expounded, and Lister's father granted membership of a hitherto unknown organization, the British Royal Society of Optics. Some of the factual blunders have been introduced into this edition through sloppy editing; others were in the original.

Debré's biography thus needs to be read with caution. There are, however, rewards. Debré writes well and his expositions of Pasteur's scientific work are clear and as full as published sources allow. He makes excellent use of the four volumes of *Correspondance* published by Pasteur's son-in-law, Louis Pasteur Vallery-Radot. His evocations of Pasteur's visit to the mansion of Napoleon III, his admission to the Académie française, and the opening ceremony of the Pasteur Institute, are inspired. The description of him as the "chemist with the microscope" is memorable. Like Geison, Debré has discovered the value of the testimony of Pasteur's nephew and collaborator, Adrien Loir. We might wish that Emile Roux had been given greater voice.

In the end, this volume is based too much on printed sources, and a too narrow range of those, to be definitive. Debré has gone beyond the saintly image that early biographers constructed, and his Pasteur was certainly a thoroughly modern scientist in the way he used the media in achieving his goals. Nor has the importance of those achievements been vitiated by other contemporary scholarship.

Rabid secret: Pasteur's rabies vaccine proved difficult to evaluate.



Art. LA RAGE est une maladie acute caractérisée par des accès de fureur, des envies de morsure, souvent accompagnées de l'horreur de l'eau, des bousses et quelquefois de convulsions à l'apogée des corps brillants et lumineux.

Nevertheless, there remains too much of the romantic in this biography of a romantic scientist who became preoccupied with his place in history. He crafted the last two decades of his life to ensure that place, which will endure. Pasteur hated nothing so much as being wrong: considering the range and imaginativeness of his research, his success on that score is perhaps the most remarkable characteristic of a remarkable scientist. But I am still not convinced I would have wanted to have had dinner with him. □

W. F. Bynum is at the Wellcome Institute for the History of Medicine, 183 Euston Road, London NW1 2BE, UK.

The face of the future

Enhancing Human Traits: Ethical and Social Implications

edited by Erik Parens

Georgetown University Press: 1998. 251 pp. \$49.95, £38.95

David Gems

As biomedical science progresses, ever more effective medical technologies are devised to treat illnesses, and this is, of course, a good thing. But how do we feel about the use of such technologies by people who are healthy to start with but who want to become *more* than healthy?

Many such enhancement technologies are already widely available. Cosmetic surgery is used for aesthetic enhancement of the body, β -blockers are taken by musicians to block the physical symptoms of performance nerves, and the antidepressant Prozac is used as an agent of what Peter Kramer has called cosmetic psychopharmacology, or alteration of personality, to make people less shy, less compulsive, more confident. We must assume that, with time, many more enhancement technologies will become available, employing surgery, genetics, pharmacology and heaven knows what else, directed in particular at cognitive function and longevity.

Are enhancement technologies a good thing? We regard self-development through education and exercise as a virtue, almost a duty. Why not pursue these ends by means of enhancement technologies? Yet to many people, enhancement technologies evoke visions of eugenics, Nazi conceptions of the superman, and Aldous Huxley's *Brave New World*.

Enhancing Human Traits represents a landmark in the discussion of these thorny issues. It is the product of a project coordinated by the philosopher Erik Parens at the Hastings Center in New York, and has brought together thinkers from various fields, including philosophy, law, sociology,

theology and women's studies, to discuss the rights and wrongs of biological enhancement.

The book contains 13 very different essays on distinct facets of this complex subject. They involve two sorts of discussion on enhancement. The first focuses on the distinction between treatment and enhancement, and concerns what doctors should and should not do, and what health-care systems should and should not provide. The second deals with the broader issue of the value of enhancement technologies *per se*. As Parens puts it, the first discussion of enhancement concerns the goals of medicine, the second the goals of society.

On closer inspection, the treatment/enhancement distinction is problematic in various ways. To a degree, what counts as illness is culturally constructed. Until recently, being gay was regarded as an illness; likewise, among the Punan Bah people of Borneo, giving birth to twins is regarded as a dreadful misfortune. Some of these nosological problems can be overcome by defining health as the full expression of species-typical function. However, as Anita Silvers points out in her contribution, this approach falls down when confronted with the functionally different, yet richly rewarding ways of life of many disabled people. Some members of the deaf community sarcastically label people who can hear as 'signing impaired'.

Of the contributing academic disciplines, the clearest analyses of the most worrying aspects of enhancement technologies come from women's studies, particularly from Margaret Little and Susan Bordo. If

current experience of cosmetic surgery is a good indication of how other enhancement technologies will come into use in future, we have plenty to worry about. Particularly alarming is the convergence of enhancement technologies and capitalism. It is only too easy to create markets for cosmetic surgery by convincing people that they possess defects that only surgery can repair. This marketing of ersatz illness has generated a Newspeak of inadequacy: a woman with little breasts becomes 'micromastic', and the dimples on her thighs become 'cellulite' — conditions requiring treatment. Legislation may be necessary to guard against such insidious, market-driven medicalizations.

A more awkward problem involves what Little calls the ethics of complicity. Consider a plastic surgeon approached by a woman who is discriminated against because of her large breasts or African appearance and who is seeking breast reduction or de-Africanization *à la* Michael Jackson. As Little points out, by agreeing to operate, the surgeon becomes complicit with harmful conceptions of normality — in this case, norms that are oppressive to women, or racist. On the other hand, the suffering of those seeking cosmetic surgery — including that resulting from oppressive standards of normality — may be very real. Little suggests that the surgeon may act ethically by performing the surgery, but only if she also works elsewhere to fight the system of unjust practices and attitudes.

The significance of enhancement technologies is also discussed by reference to ideas about what makes life meaningful and



"She's had so much done she's not even biodegradable."

WILLIAM HAMILTON/NEW YORKER

worth living. As theologian Ronald Cole-Turner points out: "Technologies such as psychopharmacology and human genetic manipulation fit very well within the broad program of Western religions and philosophy". In this context, philosopher Carl Elliott expresses concern that drugs such as Prozac could be used to 'treat' alienation or existential anxiety — or, as contributor Gerald McKenny puts it: "To relieve the human condition". To Elliott, seeing this sort of unhappiness as a psychiatric issue represents a category error, "like seeing Holy Communion as a dietary issue".

One way to see what is so disturbing about, say, a happy, smiling August Strindberg on Prozac is that he might no longer be Strindberg. Elliott refers to what the philosopher Charles Taylor has called an ethics of authenticity: the idea that our lives can only have meaning insofar as we are true to ourselves, and on this basis develop our own life project. Thus, altering one's personality by means of Prozac may potentially be an act of self-betrayal resulting in a seemingly happy, yet phoney and pointless life.

This is a fascinating, challenging and important book, and a major achievement by Parens. Although the contributions are written by academics, by the style of most of them they are obviously meant for a broader audience, and use clear and accessible language — Dan Brock's essay in particular. I predict that this book will open a debate that will play a significant role in shaping our culture in the years to come. □

David Gems is in the Galton Laboratory, Department of Biology, University College London, London NW1 2HE, UK.

Destruction and deliberation

The Earth in Turmoil: Earthquakes, Volcanoes, and Their Impact on Humankind

by Kerry Sieh and Simon LeVay

W. H. Freeman: 1998. 323 pp. \$24.95, £14.83

Claudio Vita-Finzi

Just over a century ago, the pioneering seismologist John Milne, then at the Imperial College of Engineering in Tokyo, was trying to decide whether the apparent association between earthquakes and volcanoes was a coincidence. "It would seem," he concluded, "that the two phenomena are without any direct connection, unless it be that both are different effects of a common cause."

Geology has since supplied the requisite common cause: the collision between or splitting apart of two of the plates that make up the Earth's outer shell. But the layman has never doubted that earthquakes and volcanoes go together, like plague and famine or



Awful sight: we know much about volcanoes, but are still often caught off guard when they erupt.

fire and brimstone. As Pliny and anyone else living under Mount Vesuvius knows, they often occur together and they epitomize the hostile side of nature.

Kerry Sieh and Simon LeVay in *The Earth in Turmoil* respect this fear — and its obverse, the desire to do something about the twin scourges. More than 1.5 million people have died in earthquakes this century and the 1985 eruption of the Nevado del Ruiz volcano in Colombia killed 25,000. But, with the detachment of neurologists confronted by a devastating cranial lesion, the authors also welcome eruptions and earthquakes as clues to the workings of the planet.

Their casebook is provided by the United States, which has its share of volcanoes, earthquakes and irresponsible planners, and, more to the point, superlative geoscientists and research organizations. In contrast with many of their European counterparts, these scientists and organizations make data, facilities and ideas available to the world at large with prodigal liberality.

It is not always a question of money. The geologist Brian Atwater, armed with little more than a shovel and a canoe, revealed to the complacent population of the north-western United States that it was at greater risk from major earthquakes, and the attendant sea waves and landslides, than the Californians down the coast. The local written history was too short to register a major event 300 years ago comparable in

severity to the earthquake that devastated south-central Chile in 1960 (the largest in recorded history), but sand and peat layers revealed a sequence of such events separated by between three and nine centuries.

Our knowledge of the much more notorious San Andreas fault has also benefited from historical analysis, and one of the 23 colour plates in the book could well have been swapped for a detailed trench section by Sieh. For here, too, an inadequate human chronicle has been remedied by the spade — and the mechanical digger — to reveal at least two nineteenth-century earthquakes on the scale of the 1906 San Francisco disaster.

Seismic zoning has not kept up with understanding, nor with the humility it brings. Once movement on faults was recognized as the cause rather than the result of earthquakes, the emphasis was put on mapping them and estimating the probability of when and how they would move. Sieh and LeVay have a fine time listing some of the major cities that lie within easy reach of a major active fault, among them Tokyo, Osaka, Xian, Beijing, Lima, Valparaíso, Algiers, Athens, Jerusalem, Tehran, Karachi and Kabul.

Yet many of the most destructive events have occurred on hidden structures whose existence was signalled only by the mayhem they caused. And the earthquakes of 1811–12 in New Madrid in the United States, like the 1993 Latur earthquake in India that killed 10,000 people, remain puzzling because their locations do not conform to the oversimplified versions of plate tectonics.

The situation is similar with volcanoes. We know a lot, but are caught off guard either because we ignore warnings or because eruptions are triggered by apparently unrelated events such as landslides. The forces involved are prodigious: during its 1980 eruption, Mount St Helens liberated as much energy as the Hiroshima bomb every second for nine hours. As in the study of earthquakes, there is much to be learnt from folk tales which may point to a prehistoric event for which the evidence is ephemeral. A piquant example is the eruption that is thought to have prompted two medicine men to throw themselves into Mount Mazama in Oregon, to placate the Chief of the Below World.

Scientists continue to die in their efforts to understand or predict eruptions. This book documents their achievements. It does so with eloquent pride and with greater incision than many college textbooks. One quibble: the title speaks of the Earth in turmoil. When has it been at peace? □

Claudio Vita-Finzi is in the Department of Geological Sciences, University College London, London WC1E 6BT, UK.

Antidote to angst

Wissenschaft gegen Zukunftsangst

by Hubert Markl

Hanser: 1998. 362 pp. DM45

Gerhard Neuweiler

Zukunftsangst is difficult to translate; it refers to amorphous feelings of concern and anxiety about the future. Although this book's title, *Wissenschaft gegen Zukunftsangst*, indicates its aims to show how science can overcome *Zukunftsangst*, the term is covered by only a few sentences. Perhaps 'Follow me' would be a more appropriate title for this collection of 16 public lectures given by Hubert Markl.

As a former president of the German research council and current president of the Max Planck Society, Markl is the most prominent and powerful voice of German science today. Throughout the book he relentlessly canvasses for a world guided by science. Markl claims that experience over the past three centuries has taught us that only the methods and insights of science are consistently reliable.

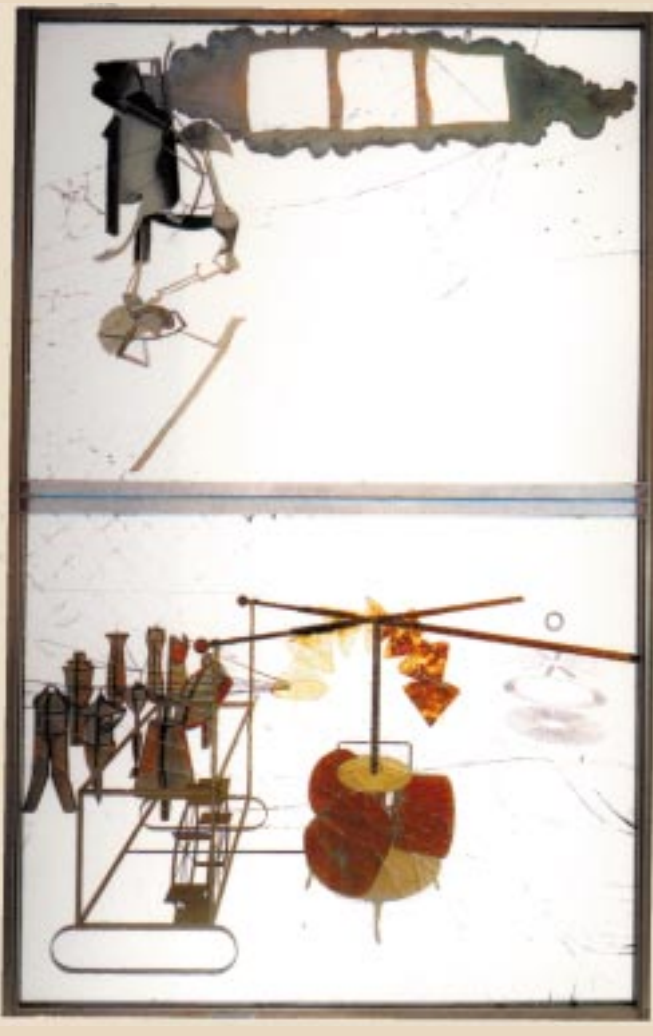
In support of this conviction, the first seven lectures present an array of facts and concepts from modern biology. We are told how man evolved from the animal kingdom through social learning and cognition, and why "there is no such thing as a free lunch" in an ecologically interconnected biosphere that will not dispense with mankind. This part of the book makes highly informative reading for physical scientists. The wealth of facts are lucidly presented. The language sparkles with colourful metaphors, inspiring analogies and references to historical events and classical literature. This is media science at its best, and these chapters should be compulsory reading for decision-makers in public affairs.

Markl identifies certain challenges that must be addressed if we are to achieve a sustainable future for humankind. Among these are control of human population size (underlined in red and repeated throughout the book); sustainable management of the biosphere; the fight against the lobbying power of special interest cartels; and support for science and technology.

Although we may agree with these aims, it is clear that progress requires brave cuts into a thick and dangerous jungle of competing interests. But, surprisingly, the book offers a plain highway to a promising future: keep to rationality, follow the signs of science and you will travel in the right direction and finally arrive at your destination. On this road there seems no place for philosophy, psychology and sociology, or for any science not brand-

A 'playful physics'

Marcel Duchamp wrote hundreds of preparatory notes for his main work: two panes of glass entitled "The Bride Stripped Bare by Her Bachelors, Even". He intended these notes — many of which discuss science in the early twentieth century — to accompany the glass "somewhat like a Sears Roebuck catalogue" and to equal it in importance. Later in his life he denied any serious interest in science and technology. From *Duchamp in Context: Science and Technology in the Large Glass and Related Works* by Linda Dalrymple Henderson (Princeton University Press, \$85, £60).



ed by the fire of experiment.

It is at this juncture that the book loses much of its power. Although the author acknowledges that mankind is also driven by irrational motivations, and that there are the arts, poetry and other creative human activities, these are not included in his framework for a future world. All the lectures remain strictly within the natural sciences with their spotless paths weeded by infallible logic.

This is why the second part of the book, which addresses interrelations between science and society, and power and politics, is less inspiring and informative. There is a lecture on science and power, a field in which Markl has been involved for decades. A discussion of how science deals with the political and economic power on which it depends would have been fascinating. The issue is not discussed, however. The gardener of science only passes on to these powers a list of fertilizers needed to keep the garden clean, flourishing and productive.

From an ethological viewpoint, *Zukunftsangst* is described as one of the costs of being human. "When this human trait becomes irrational (*sic*) and no longer differentiates between the possibility and

probability of imagined dangers, it provokes a pathological epidemic." The author does not discuss its nature nor why it has reached epidemic proportions in Germany. There is no reflection on aspects of German history that might explain the specific virulence of *Zukunftsangst* there, especially among intellectual groups anxious to prevent any disaster to humanity emanating again from this country.

In the book, *Zukunftsangst* is used merely as a punch-bag to demonstrate the power and strength of science. Although I agree with the author that the only effective antidote to an epidemic of *Zukunftsangst* is knowledge, when a fighter does not take his opponent seriously the fight loses its impact.

Consequently, when the reader closes this enlightening book, he might behave like the fabled Bavarian peasant, who leaves church moved by an inspiring sermon and goes straight to the nearest inn for a stein of beer over which to plan how to sell his ailing cow to an unsuspecting buyer. □

Gerhard Neuweiler is at the Zoologisches Institut, Ludwig-Maximilians-Universität, D-80021, Munich, Germany.

The large-scale smoothness of the Universe

Kelvin K. S. Wu, Ofer Lahav & Martin J. Rees

The Universe is inhomogeneous—and essentially fractal—on the scale of galaxies and clusters of galaxies, but most cosmologists believe that on larger scales it becomes isotropic and homogeneous: this is the ‘cosmological principle’. This principle was first adopted when observational cosmology was in its infancy, and was then little more than a conjecture. The data now available offer a quantitative picture of the gradual transition from small-scale fractal behaviour to large-scale homogeneity.

A fractal is a distribution or shape that is not homogeneous (in general), but possesses the property that each part is a simulacrum of the whole. In other words it ‘looks’ the same on all scales. Fractals abound in nature: for example, the coastline of a small peninsula drawn on paper could equally well depict a large continent; and fractals have long been studied as manifestations of scaling laws in solid-state physics^{1,2}. The clustering of galaxies (see Fig. 1) lends itself to a fractal description^{1–7}, because the clumpiness prevails over a wide range of scales. The big question for cosmology is whether the distribution of matter continues to be a simple fractal beyond the scale of clusters of galaxies: if it does, then the cosmological principle is invalid.

Past attempts to determine the distribution of matter have confronted two important obstacles. Most of the mass in the Universe is dark; we can detect it only by its gravitational effect on objects that we can see, and it is still unclear how to relate the distributions of light and mass. In particular, we have not known how to match the clustering of galaxies with the cosmic microwave background (CMB) anisotropies, which tell us about the mass fluctuations in the early Universe. In addition, little has been known about fluctuations in the distribution of matter on scales intermediate between those of local galaxy surveys ($\leq 100 h^{-1}$ Mpc, where h is the Hubble constant in units of $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$) and scales probed by the Cosmic Background Explorer (COBE) satellite ($\geq 1,000 h^{-1}$ Mpc).

Recent observations of radio galaxies and the X-ray background (XRB, which probably arises from distant active galactic nuclei), at median redshift $\bar{z} \approx 1$, can now effectively probe these intermediate scales. Moreover, experiments to measure fluctuations in the microwave background on smaller angular scales than those probed by COBE are also helping to bridge the gap. Future surveys of more than one million galaxies will probe to a median redshift $\bar{z} \approx 0.1$ (which roughly corresponds to a co-moving distance of $\sim 300 h^{-1}$ Mpc). Current data, however, already strongly constrain any non-uniformities in the galaxy distribution (as well as the overall mass distribution) on scales $\geq 300 h^{-1}$ Mpc.

In the language of fractals (see Box 1), fractal dimensions are used to characterize the degree of clustering; these generalize our intuitive concepts of dimension and can take any positive value less than or equal to 3. On scales below $\sim 10 h^{-1}$ Mpc, galaxies are distributed with the correlation dimension⁸ $D_2 = 1.2\text{--}2.2$ (Box 1 and Table 1). But we show below that on large scales D_2 is very close to the homogeneous value of 3. Any quantitative discussion of the large-scale structure in the Universe in fact depends on the unknown cosmological parameters (defined only for a homogeneous and isotropic universe): the density parameter Ω , the cosmological constant Λ and the Hubble constant H_0 . For simplicity we present the observational results interpreted for the Einstein-de Sitter model ($\Omega = 1$ and $\Lambda = 0$), but the main conclusions are not altered for other models.

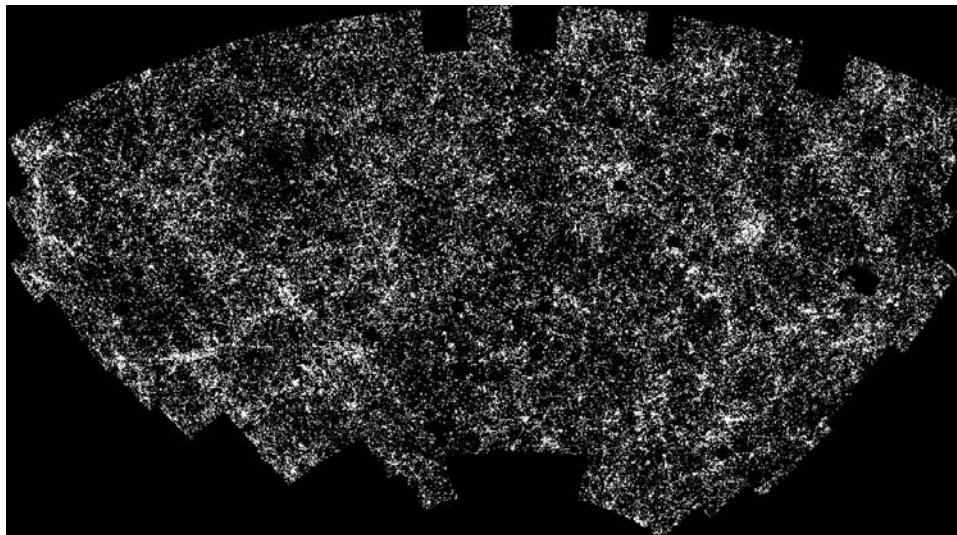


Figure 1 The distribution of 2 million galaxies with blue magnitude $17 \leq b_j \leq 20.5$ shown in an equal-area projection centred on the south Galactic pole. The data are from scans over a contiguous area of 4,300 square degrees using the Automatic Plate Measuring (APM) machine¹⁰. The small empty patches in the

map are regions excluded around bright stars, nearby dwarf galaxies, globular clusters and regions removed for calibration purposes. Although projection and superposition effects tend to wash out three-dimensional structures, the pattern is seen to be non-uniform, with clusters, filaments and voids.

Box 1 The fractal dimension

If we count, for each galaxy, the number of galaxies within a distance R from it, and call the average number obtained $N(<R)$, then the distribution is said to be a fractal of correlation dimension D_2 if $N(<R) \propto R^{D_2}$. Of course D_2 may be 3, in which case the distribution is homogeneous rather than fractal. In the pure fractal model, this power law holds for all scales of R , whereas in a hybrid model it holds for R less than some scale, above which D_2 increases towards 3 to accommodate the cosmological principle. To allow for varying D_2 one often writes:

$$D_2 = \frac{d(\ln N(<R))}{d(\ln R)}$$

Using the above, the fractal proponents^{4,5} have estimated $D_2 \approx 2$ for all scales up to $\sim 1,000 h^{-1}$ Mpc, whereas other groups^{6,18,66–70} have obtained, in general, scale-dependent values as listed in Table 1.

These measurements can be directly compared with the popular cold dark matter (CDM) models of density fluctuations, which predict the increase of D_2 with R . If we now assume homogeneity on large scales, then the mean density $\bar{n}$ and the correlation function $\xi(r)$ can be defined, and:

$$N(<R) = \frac{4\pi}{3} R^3 \bar{n} + 4\pi \bar{n} \int_0^R dr r^2 \xi(r)$$

for a flat universe with $\Omega = 1$. Hence we have a direct mapping between ξ and D_2 . If we choose a power-law form for $\xi(r)$ (equation (1) in Box 2), then it follows⁸⁸ that $D_2 = 3 - \gamma$ if $\xi \gg 1$. If $\xi(r) = 0$, we obtain $D_2 = 3$. The CDM models give us the correlation function $\xi(r)$ on scales greater than $\sim 10 h^{-1}$ Mpc, where we do not need to worry about nonlinear gravitational effects. The function $N(<R)$ can then be calculated from these correlations. The predicted runs of $D_2(R)$ from three different CDM models are given in Fig. 4. They may differ somewhat but they all show the same qualitative behaviour: above $30 h^{-1}$ Mpc we should be able to measure dimensions close to 3, not 2. Above $100 h^{-1}$ Mpc, they become indistinguishably close to 3. They also illustrate that it is inappropriate to quote a single crossover scale to homogeneity, for the crossover is gradual. Here we have described but one statistical fractal measure, D_2 , out of a much larger set known as the 'multifractal spectrum', which is a useful tool for the statistical description of redshift surveys⁸⁹.

Local galaxies are strongly clustered

The clumpiness of matter in the Universe was initially studied by measuring the clustering of bright galaxies^{8,9}. Figure 1 shows 2 million optically selected galaxies¹⁰ projected on the sky. The distribution is evidently not uniform: galaxies are arranged in clusters and 'filaments', and avoid certain regions termed 'voids'. Figure 2 shows data from the largest redshift survey to date, Las Campanas¹¹, which illustrates (insofar as the redshift of each galaxy indicates its distance) the three-dimensional clustering of galaxies. Although clustering is seen on scales of tens of megaparsecs, on larger scales the distribution seems more homogeneous.

It is well established that the probability of finding a galaxy

$\sim 5 h^{-1}$ Mpc away from another galaxy is twice the probability expected in a uniform distribution (Box 2). The clustering of optical galaxies^{12,13} is illustrated in Fig. 3, which shows $\langle (\delta\rho/\rho)^2 \rangle$, where ρ is the density and $\delta\rho$ is the fluctuation, as a function of characteristic length scale λ . The solid and dashed lines correspond to two variants of the cold dark matter (CDM) model for mass density fluctuations¹⁴, which is widely used as a 'template' for comparison with data. We see that the fluctuations drop monotonically with scale (although not as a pure power-law). On a scale of $\lambda \approx 100 h^{-1}$ Mpc, the root-mean-square (r.m.s.) fluctuation is only 10%. This is the crucial evidence that on larger scales the fluctuations become negligible.

It is most unlikely that luminous galaxies trace perfectly the mass distribution. Galaxies can only form in dense regions, and their formation may be affected by other physical conditions and local environment: the clustering of galaxies is therefore likely to be 'biased' relative to the mass fluctuations (Box 2). Indeed, the galaxy distribution could in principle display conspicuous features on very large scales even if the mass did not—for instance, a long cosmic string could 'seed' galaxy formation in its wake. So the galaxies could be arrayed in a fractal structure, even if the mass distribution is non-fractal. It is important therefore to understand the biasing in order to match the fluctuations in galaxies to the fluctuations in mass.

Other (biased) probes at large distances are clusters of galaxies, as selected optically by Abell¹⁵ or by X-ray surveys¹⁶. These surveys typically probe out to redshift $z \approx 0.1$. Several studies^{17,18} suggest that on scales of $\sim 600 h^{-1}$ Mpc, the distribution of Abell clusters is homogeneous.

A more controversial result on the distribution of galaxies suggests a 'characteristic scale' of clustering of $\sim 128 h^{-1}$ Mpc (refs 19, 20). It is not clear yet if this feature is real, or just due to small-number statistics or survey geometry²¹. Einasto *et al.*²² have suggested that Abell clusters lie in a quasiregular three-dimensional network of superclusters and voids, with regions of high density separated by about $120 h^{-1}$ Mpc. The reality of such a 'periodicity' in galaxy clustering should soon be revisited by two new large redshift surveys. The American–Japanese Sloan Digital Sky Survey (SDSS) should yield redshifts for about 1 million galaxies, and the Anglo–Australian '2 degree Field' (2dF) survey should produce redshifts for 250,000 galaxies (both with median redshift of $\bar{z} \approx 0.1$). These big galaxy surveys should provide good statistics on scales larger than $\sim 100 h^{-1}$ Mpc.

Tracers at high redshift

We need to observe beyond $z = 0.1$ in order to sample a big enough volume to probe clustering on scales above $\sim 300 h^{-1}$ Mpc, and to fill the gap between scales probed by galaxy surveys and the scales probed by COBE. However, this then introduces the extra complication that we cannot interpret the data without taking account of how the clustering evolves with time, and also possible cosmic evolutionary effects in the brightness of objects. The XRB and radio

Table 1 Estimates of the fractal correlation dimensions

	Sample	R (h^{-1} Mpc)	D_2
Guzzo <i>et al.</i> ⁶⁶	Perseus-Pisces	10–3.5	1.2
		3.5–20	2.2
Martinez and Coles ⁶⁷	QDOT	1.0–10	2.25
		10–50	2.77
Lemson and Sanders ⁶⁸	CfA	1.0–30	2.0
Martínez <i>et al.</i> ⁷⁰	Stromlo-APM	30–60	2.7–2.9
Scaramella <i>et al.</i> ¹⁸	ESO Slice Project	300–400	2.93
X-Ray background		~ 500	$3 - D_2 = 10^{-4}$, with $\sigma_8 = 2$, $\Gamma = 0.5$
COBE 4-year normalization		$\sim 1,000$	$3 - D_2 = 2 \times 10^{-6}$, with $\sigma_8 = 1.4$, $\Gamma = 0.5$

Some estimates of the fractal correlation dimension D_2 obtained from galaxy surveys, showing a general increase with scale R , as defined in Box 1. Scaramella *et al.*¹⁸ analysed a number of subsamples with different methods, from which we chose one of their largest. All their results that include necessary 'k-corrections', which account for the effect of galaxy spectra being redshifted relative to the observer's pass band, are consistent with $D_2 = 3$ within their errors. Also given are estimates of D_2 from the X-ray background (XRB) and the cosmic microwave background (CMB), obtained by normalizing a standard cold dark matter (CDM) model to match measured anisotropy results. Such models are characterized by the value of σ_8 , the r.m.s. fluctuation on $8 h^{-1}$ Mpc scale, and a so-called shape parameter Γ . Unlike the other measurements in this table, the CMB probes directly the fluctuations in mass.

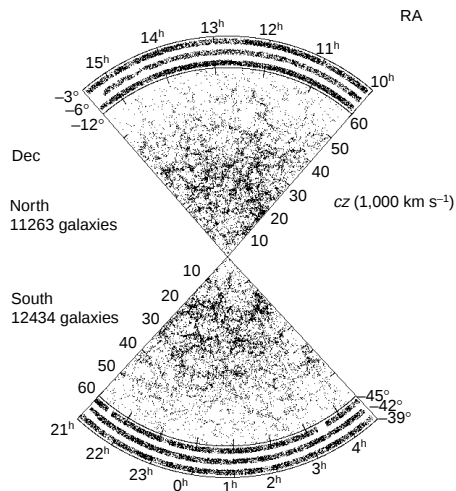


Figure 2 The redshift distribution of more than 20,000 galaxies, from the Las Campanas redshift survey¹¹. The plot shows the superposition of three slices in the north Galactic cap, and likewise for the south Galactic cap, plotted in redshift versus angular coordinates (RA, right ascension; dec., declination). Clustering of galaxies is seen on scales smaller than $\sim 30 h^{-1}$ Mpc, but on larger scales the distribution approaches homogeneity. We note that the diluted density of galaxies at higher redshifts is an artefact, due to the selection of galaxies by their apparent flux.

sources are tracers of galaxies (or at least of that subset which is active) out to median redshift $\bar{z} \approx 1$. Figure 3 shows that the limits they set to large-scale inhomogeneities ($> 100 h^{-1}$ Mpc) are less stringent than those implied by CMB measurements; but they provide independent constraints, as they sample 'luminous' objects rather than the total mass. Other possible high-redshift tracers are quasars and clusters of galaxies.

Radio galaxies. Radio sources in surveys have typical median redshift $\bar{z} \approx 1$, and hence are useful probes of clustering at high redshift. Earlier studies²³ claimed that the distribution of radio sources supports the cosmological principle. However, their wide range in intrinsic luminosities would dilute any clustering when projected on the sky. Analyses of new deep radio surveys²⁴ suggest that radio sources are clustered at least as strongly as local optical galaxies^{25–31}. Nevertheless, on very large scales the distribution of radio sources seems nearly isotropic. The measured quadrupole offers a crude estimate³² of the fluctuations on scales $\lambda \approx 600 h^{-1}$ Mpc. The derived amplitudes are shown in Fig. 3 for the two assumed CDM models. Given the problems of catalogue matching and shot-noise, these points should be interpreted as significant 'upper limits', not as detections.

The X-Ray background (XRB). Although discovered in 1962, the origin of the XRB is still controversial, but its sources, whatever they turn out to be, are likely to be at high redshift^{33,34}.

The XRB is a unique probe of fluctuations on intermediate scales between those of local galaxy surveys and COBE^{35–41}, although the interpretation of the results depends on the nature of the X-ray sources and their evolution. The r.m.s. dipole and smaller-scale fluctuations over the sky can be predicted³⁸ in the framework of gravitational instability and assumptions about the distribution of the X-ray sources with redshift. By comparing these predictions with data from HEAO⁴², it is possible to estimate the amplitude of fluctuations for an assumed shape of the density fluctuations (for example, the CDM model). Figure 3 shows the amplitude of fluctuations derived at the effective scale $\lambda \approx 600 h^{-1}$ Mpc probed by the XRB. Assuming a specific epoch-dependent biasing scheme⁴³ (Box 2) and a range of models of evolution of X-ray sources and clustering, the present-epoch density fluctuations of the X-ray sources is found to be no more than twice the amplitude of fluctuations in mass⁴².

Box 2 Quantitative measures of galaxy clustering

One popular measure of galaxy clustering is the two-point correlation function⁸. This is defined as the excess probability, relative to a random distribution, of finding a galaxy at a distance r from another galaxy. It is now well established that on scales smaller than $\sim 10 h^{-1}$ Mpc it has roughly the form:

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma} \quad (1)$$

For optically selected galaxies, $\gamma = 1.8$ and $r_0 \approx 5 h^{-1}$ Mpc; for galaxies observed in the infrared with the IRAS satellite, which include spiral galaxies but under-represent ellipticals, $r_0 \approx 4 h^{-1}$ Mpc with a somewhat shallower slope⁷⁵. The clustering of galaxy clusters (as selected optically by Abell¹⁵ or by X-ray surveys) obeys a similar law⁷⁶ but with a much stronger clustering amplitude, $r_0 \approx (15\text{--}20) h^{-1}$ Mpc. The Fourier transform of the correlation function $\xi(r)$ is the power spectrum $P(k)$ (where k is the wavenumber), which corresponds to the square of Fourier coefficients of the fluctuations. The r.m.s. fluctuations (Fig. 3) can be written as $\langle(\delta\rho/\rho)^2\rangle \propto k^3 P(k)$.

It is not likely that the fluctuations in the density of particular galaxy types are exactly the same as the fluctuations in mass. The simplest assumption, which has been widely adopted, is that the galaxy and mass density fluctuations (δ_g and δ_m , respectively) at any point $\mathbf{x}$ are related by;

$$\delta_g(\mathbf{x}) = b \delta_m(\mathbf{x}) \quad (2)$$

where b is the 'bias parameter'. Usually $b > 1$, which implies that the galaxies are more clustered than the mass distribution. By modelling galaxies as peaks of the underlying mass distribution and using an argument analogous to that which explains why the highest ocean waves come in groups, Kaiser⁷⁷ showed that in the linear approximation the correlation function of galaxies (ξ_{gg}) is related to the mass correlation function (ξ_{mm}) by;

$$\xi_{gg}(r) = b^2 \xi_{mm}(r) \quad (3)$$

where r is the separation between galaxies or mass elements. We note that although equation (3) does follow from equation (2), it is more general and does not imply equation (2). Various theoretical and observational considerations suggest that $b \approx 1\text{--}2$.

Biasing must certainly be more complicated than equations (2) and (3): indeed, clustering is not the same for galaxies of different morphologies. For example, elliptical galaxies are more strongly clustered than spiral galaxies on scales $\leq 10 h^{-1}$ Mpc (refs 78–80). The appropriate value of b may depend on scale, as well as on the local overdensity. Furthermore, it is not clear *a priori* that δ_g is just a function of δ_m . The efficiency of galaxy formation could in principle be modulated by some large-scale environmental effects (for example, heating by early quasars, or the proximity of a cosmic string) which are uncorrelated with δ_m . Biasing might therefore be non-local, nonlinear, stochastic and epoch-dependent^{43,45,81–87}.

Quasars, high-redshift galaxies and Lyman- α clouds. Until the mid-1990s, quasars—hyperactive galactic nuclei—were the only objects luminous enough to be identified in substantial numbers at redshifts $z > 2$. It is still unclear how their clustering evolves with redshift⁴⁴, but they appear no more clustered than extragalactic radio sources (which are a related population). The advent of 10-metre-class telescopes now allows the detection of galaxies out to equally large redshifts. Several hundred galaxies with $z \geq 3$ have already been detected, and they display about the same level of clustering as nearby galaxies⁴⁵. As gravitational effects enhance density contrasts during cosmic expansion, one might at first sight have expected weaker clustering of galaxies at earlier times. However, the luminous galaxies that have already formed at the epoch corresponding to $z = 3$ belong to an exceptional subset associated with unusually high density peaks, which display enhanced clustering (Box 2). When this is taken into account, the data are compatible with CDM models normalized to match the degree of clustering at low redshifts⁴⁵.

The rich absorption-line spectra of high-redshift quasars offer a probe for the distribution of intervening material. In particular, the 'Lyman- α forest' in quasar spectra reveals the distribution of diffuse gas clouds along the line of sight. There are no large 'clearings' in the Lyman- α forest: indeed the relevant clouds seem even more smoothly distributed than galaxies⁴⁶. Moreover, the spectra of different quasars indicate that the clouds have similar properties along all lines of sight. Relating the spacings of these clouds to the overall mass distribution (or even to the galaxy distribution) is not straightforward. However, if the clouds manifested scale-free fractal-like structure it would be remarkable if this structure were not plainly apparent in quasar absorption spectra⁷.

Direct probes of mass distribution

We have emphasized that the luminous objects selected by surveys may not trace the total mass. There are, however, at least three independent probes of inhomogeneities in the gravitational field induced by the total mass fluctuations: lensing, the CMB, and peculiar velocities. Gravitational lensing—the distortion of distant galaxy images by intervening potential wells—cannot, as yet, constrain the mass fluctuations on scales larger than $20 h^{-1}$ Mpc or so^{47,48}. Here we focus on the other two probes of mass fluctuations, which show good consistency with the picture that

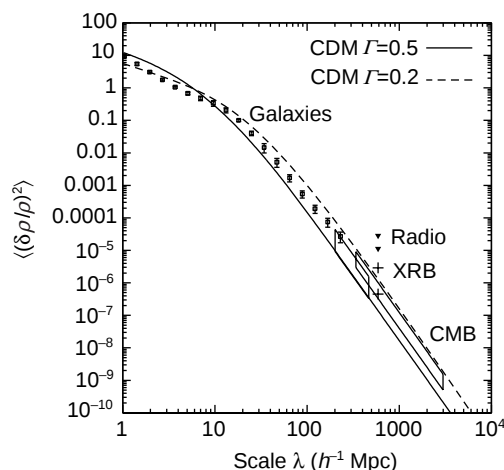


Figure 3 A compilation of density fluctuations on different scales from various observations. Shown are data from a galaxy survey, deep radio surveys, the X-ray background (XRB) and cosmic microwave background (CMB) experiments. The measurements are compared with two popular cold dark matter (CDM) models. The figure shows mean-square density fluctuations $\langle(\delta\rho/\rho)^2\rangle$. The solid and dashed lines correspond respectively to the standard CDM power spectrum (with shape parameter $\Gamma = 0.5$) and a 'low-density' CDM power spectrum (with $\Gamma = 0.2$). Both models are normalized such that the r.m.s. fluctuation within $8 h^{-1}$ Mpc spheres is $\sigma_{8,M} = 1$. The open squares at small scales are estimates of the power spectrum from three-dimensional inversion of the angular APM galaxy catalogue^{12,13}. The elongated 'boxes' at large scales represent the COBE 4-yr (refs 57, 90, 91) (on the right) and Tenerife⁹² (on the left) CMB measurements. The filled triangles represent constraints from the quadrupole moment of the distribution of radio sources³². This quadrupole measurement probes fluctuations on scale $\lambda_* \approx 600 h^{-1}$ Mpc. The top and bottom filled triangles are upper limits of the amplitude of the power spectrum at λ_* , assuming CDM power spectra with shape parameters $\Gamma = 0.2$ and 0.5 , respectively, and an Einstein-de Sitter universe. The crosses represent constraints from the XRB HEAO1 quadrupole^{39,42}. Assuming evolution, clustering and epoch-dependent biasing prescriptions, this XRB quadrupole measurement probes fluctuations on scale $\lambda_* \approx 600 h^{-1}$ Mpc, very similar to the scale probed by the radio sources. The top and bottom crosses are estimates of the amplitude of the power spectrum at λ_* , assuming CDM power spectra with shape parameters $\Gamma = 0.2$ and 0.5 respectively, and an Einstein-de Sitter universe. The fractional error on the XRB amplitudes (due to the shot noise of the X-ray sources) is $\sim 30\%$.

amplitudes are significant on small scales but are tiny on the very large scales.

The cosmic microwave background (CMB). The CMB is well described by a black-body radiation spectrum at a temperature of 2.73 K, hence providing crucial evidence for the hot Big Bang model. This sea of radiation is highly isotropic, the main anisotropy being due to the motion of our Galaxy (the Milky Way) at 600 km s^{-1} relative to the CMB. This motion is remarkably well reconstructed in both amplitude and direction by summing up the forces due to masses represented by galaxies^{49,50} at distances nearer than $\sim 100 h^{-1}$ Mpc. The dipole anisotropy in the distribution of nearby supernovae also indicates that most of the Galaxy's motion arises from local inhomogeneities⁵¹.

Apart from the dipole anisotropy, the other main anisotropies in the CMB radiation are imprints at the last scattering surface at redshift $z \approx 1,000$ where the primordial plasma recombined. In 1992, the COBE satellite detected fluctuations at the level of 10^{-5} on scales of 10° ; this corresponds to a present-epoch length-scale of $\sim 1,000 h^{-1}$ Mpc (Fig. 3). These tiny CMB fluctuations are attributed to 'metric' or 'curvature' fluctuations⁵² of this order in a universe which has an approximately Friedmann–Robertson–Walker (FRW) metric of space-time (corresponding to a homogeneous and isotropic universe).

On scales smaller than $1\text{--}2^\circ$, extra contributions to the temperature fluctuations arise from motions of the plasma induced by the metric fluctuations. This interaction between plasma and gravity translates to peaks in the angular power-spectrum of the temperature fluctuations^{53–56}. The angular scales of these peaks correspond to linear scales of a few hundred megaparsecs today. Several ground-based or balloon experiments are mapping small areas of sky with resolutions of $10'$ to 2° ; early in the next century, the MAP and Planck satellites should offer all-sky coverage with this resolution. The position and height of the peaks can be used to determine the cosmological parameters with very high precision^{55–58}. Moreover, the deep galaxy maps soon to be produced by the SDSS and 2dF surveys may reveal fluctuations in the galaxy distribution, at the level of a few per cent, on scales of several hundred megaparsecs; these could then be correlated with the peaks and troughs in the CMB angular fluctuation spectrum. Fluctuations probed by radio sources and the XRB on such scales can be similarly compared to the CMB (Fig. 3).

Could large-amplitude inhomogeneities along the line of sight wash out large intrinsic fluctuations in the CMB and make it look very smooth? In fact, the general effect would be merely to distort the angular distribution of the fluctuations rather than to homogenize the temperature map^{59,60}.

Peculiar velocities. Peculiar velocities (like the 600 km s^{-1} motion of our Galaxy described above) are deviations from the recession velocity that would be expected due to the smooth expansion of the Universe. The gross features of the local peculiar velocity field inferred in this way correlate well with overdensities in galaxy distribution, for example, the Virgo cluster and the Great Attractor^{61–63}, although in some regions the agreement is not perfect, perhaps due to systematic measurement errors.

Unfortunately, the distance measurement errors increase with distance, so that peculiar velocities can only be measured reliably⁶⁴ out to distances of $\sim 20 h^{-1}$ Mpc. Lauer and Postman⁶⁵ claimed that a sample of Abell clusters out to $150 h^{-1}$ Mpc is moving at $\sim 700 \text{ km s}^{-1}$ with respect to the CMB, suggesting that the CMB dipole (caused by such relative motion) is generated largely by mass concentrations beyond $\sim 100 h^{-1}$ Mpc. However, most other studies suggest bulk flows on smaller scales.

The agreement between the CMB dipole and the dipole anisotropy of relatively nearby galaxies argues in favour of large-scale homogeneity: a given overdensity $\delta\rho$ on a scale λ produces a peculiar velocity proportional to λ (in the linear regime), so a bigger peculiar velocity would be induced by larger-scale features unless $\delta\rho$ decreased as steeply as $\propto 1/\lambda$.

Large-scale structure

Although the Universe has a fractal structure on small scales, the foregoing discussion suggests that the mass distribution approaches homogeneity on large scales; that is, the fractal dimension must make a transition to $D_2 \approx 3$ (Box 1 and Fig. 4). Luminous matter does not necessarily trace mass: the galaxy distribution could in principle have been highly irregular on large scales if, for instance, galaxy formation were seeded by topological defects (for example, strings) uncorrelated with the large-scale mass distribution.

In recent years, Pietronero and co-workers^{3–5} have strongly advocated that the scale of homogeneity has not been detected even in the deepest redshift surveys. Most analyses of density fluctuations had assumed large-scale homogeneity, but these authors applied methods that made no such assumptions and argued that the fractal behaviour extended to the largest scales probed ($\sim 1,000 h^{-1}$ Mpc), with $D_2 \approx 2$. However, CDM models of density fluctuations (which fit reasonably well the available observational data) predict that at scales above $\sim 10 h^{-1}$ Mpc one should begin to detect values of D_2 greater than 2, with $D_2 \approx 3$ on scales larger than $\sim 100 h^{-1}$ Mpc (Box 1 and Fig. 4), in conflict with Pietronero's claim.

Several authors have therefore made further analyses of galaxy distributions using fractal algorithms^{18,66–70}. All of them obtained results which were consistent with standard models of density fluctuations, and all appeared to detect an approach to homogeneity on the largest scales analysed by them. We list some results for D_2 in Table 1. Although the statistics are still poor (and the quoted results do not all agree with each other), one can already see a steady increase towards $D_2 = 3$. In particular, the results do not support a constant D_2 for all scales, and the latest results by Scaramella *et al.*¹⁸ provide one of the closest fractal measurements yet to homogeneity. Scaramella *et al.* pointed out the importance of appropriate corrections to the observed flux from high-redshift galaxies to account for redshifting of their spectrum across the observer's waveband. These spectral band corrections are crucial to the interpretation of high-redshift photometry and should not be left out, approximate though they may be.

This debate has made a positive contribution by highlighting some technical issues. For example, it was correctly pointed out³ that if a survey is too small, then one cannot define the mean density (as the galaxies do form a fractal on small scales) and hence related tools such as correlation functions can be misleading. In analysing

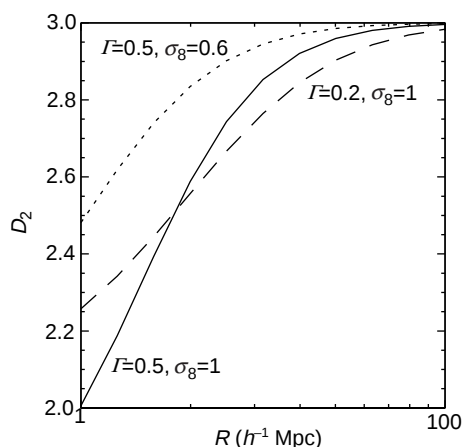


Figure 4 The fractal correlation dimension D_2 versus length scale R . The analysis assumes CDM models of power spectra with shape and normalization parameters ($\Gamma = 0.5$, $\sigma_8 = 0.6$), ($\Gamma = 0.5$, $\sigma_8 = 1.0$) and ($\Gamma = 0.2$, $\sigma_8 = 1.0$). Regardless of model, they all show the same qualitative behaviour of increasing D_2 with R , becoming vanishingly close to 3 for $R > 100 h^{-1}$ Mpc. A pure fractal model^{3–5} corresponds to the horizontal axis $D_2 = 2$, in conflict with CDM models and the data presented in Table 1.

local volumes within $\sim 30 h^{-1}$ Mpc, the fractal nature of clustering implies that one has to exercise caution when using statistical methods which assume homogeneity. On the other hand, the proponents of fractals have not helped their cause by using very incomplete and inhomogeneous samples. Detailed arguments for⁵ and against^{7,71,72} fractals on large scales have been given. The continued application of fractal algorithms to larger and deeper surveys should definitively resolve the matter.

The dependence of both the number counts and the angular correlation function of galaxies on apparent luminosity are strong evidence against a pure fractal universe⁹. Furthermore, because properties of a pure fractal are independent of scale, the projected galaxy distribution in shells of increasing size should look the same. This is strongly in conflict with the observations⁷, which show decreasing clumpiness in large shells. However, we note that visual impression alone cannot indicate the closeness to isotropy.

We consider that the agreement of XRB and CMB fluctuations with the popular CDM models—within the framework of a homogeneous universe—argues strongly against a pure fractal distribution for mass, or even for galaxies. (We remind the reader that the XRB traces galaxies, whereas the CMB traces mass.) Although direct estimates of D_2 are not possible on the scales probed by the XRB and the CMB, we can calculate their values by using CDM models normalized with the XRB and CMB as described above. The resulting values are extremely close to 3 and are given in the lower part of Table 1. (They are even tighter than the constraint $3 - D_2 \leq 0.001$ obtained by Peebles⁹ from the XRB using a different argument.) Can we go further? Isotropy does not imply homogeneity, but the near-isotropy of the CMB can be combined with the copernican principle that we are not in a preferred position. All observers would then measure the same near-isotropy, and it can be proved that the Universe must then be very well approximated by the FRW metric^{59,73}.

Although we reject the pure fractal model where D_2 is a constant on all scales, it remains worthwhile to explore further models that do not, *a priori*, assume gaussian fluctuations in a homogeneous background.

A common assumption is that the *metric* fluctuations—the deviations of the early universe from an exact FRW model—had the same amplitude on every scale. These fluctuations can be considered as scale-independent irregularities in the gravitational potential: ‘inflationary’ models suggest how they might have arisen, and the CMB data suggest an amplitude of about 10^{-5} . Such fluctuations would evolve into nonlinear clusters whose gravitational binding energy is $\sim 10^{-5}$ times their rest-mass energy: this corresponds to a virial velocity of $\sim 10^{-2.5}$ times the velocity of light, or $\sim 1,000 \text{ km s}^{-1}$. We do indeed observe such clusters in the present Universe. But we would expect a natural upper limit to the scale of nonlinear structures. Metric perturbations of $\sim 10^{-5}$ are too weak to have halted the expansion of regions whose boundaries are expanding, due to the Hubble flow, at much more than $1,000 \text{ km s}^{-1}$: they would merely have induced slight perturbations in density, with amplitude inversely proportional to the square of the scale. If the present-day *density* fluctuations in the cosmic mass distribution were scale-independent, the associated potential (or metric) fluctuations would increase as the square of the scale. It seems clear that we do not live in such a universe: it is, on the contrary, the initial fluctuations in the *metric* (rather than in the mass density) that are scale-independent. This would not formally preclude a simple fractal distribution of luminous objects, if these did not ‘track’ the overall mass distribution; we have, however, shown that galaxies, radio sources and the sources of the XRB have a distribution that becomes smoother as we average over larger scales. Of course, we still cannot formally exclude a pre-copernican model universe, such that the isotropy around us is atypical of what would be measured by hypothetical observers elsewhere⁷⁴. But, leaving such matters aside, there is a well defined sense in which our Universe is

homogeneous on the largest accessible scales; neither its mass distribution, nor that of the galaxies, resembles a pure fractal. Cosmological parameters such as Ω therefore have a well defined meaning—indeed these considerations tell us over what volume we need to average in order to determine them with any specified level of precision. $\square$

K.K.S.W. and M.J.R. are at the Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK. O.L. is at the Racah Institute of Physics, The Hebrew University, Jerusalem 91904, Israel and at the Institute of Astronomy, University of Cambridge, Cambridge, UK.

- Mandelbrot, B. B. *The Fractal Geometry of Nature* (Freeman, New York, 1983).
- Mandelbrot, B. B. in *Current Topics in Astrofundamental Physics: Primordial Cosmology* (eds Sánchez, N. & Zichichi, A.) 583–602 (Kluwer, Dordrecht, 1998).
- Coleman, P. H., Pietronero, L. & Sanders, R. H. Absence of any characteristic correlation length in the CfA galaxy catalogue. *Astron. Astrophys.* **200**, L32–L34 (1988).
- Pietronero, L., Montuori, M. & Sylos-Labini, F. On the fractal structure of the visible universe. In *Critical Dialogues in Cosmology* (ed. Turok, N.) 24–49 (World Scientific, Singapore, 1997).
- Sylos Labini, F., Montuori, M. & Pietronero, L. Scale-invariance of galaxy clustering. *Phys. Rep.* **293**, 61–226 (1998).
- Borgani, S. Scaling in the universe. *Phys. Rep.* **251**, 2–152 (1995).
- Davis, M. Is the universe homogeneous on large scales? In *Critical Dialogues in Cosmology* (ed. Turok, N.) 13–23 (World Scientific, Singapore, 1997).
- Peebles, P. J. E. *The Large-Scale Structure of the Universe* (Princeton Univ. Press, 1980).
- Peebles, P. J. E. *Principles of Physical Cosmology* (Princeton Univ. Press, 1993).
- Maddox, S. J., Efstathiou, G., Sutherland, W. J. & Loveday, J. Galaxy correlations on large scales. *Mon. Not. R. Astron. Soc.* **242**, 43P–47P (1990).
- Shectman, S. A. et al. The Las Campanas Redshift Survey. *Astrophys. J.* **470**, 172–188 (1996).
- Baugh, C. M. & Efstathiou, G. The three-dimensional power spectrum measured from the APM Galaxy Survey-1. Use of the angular correlation function. *Mon. Not. R. Astron. Soc.* **265**, 145–156 (1993).
- Baugh, C. M. & Efstathiou, G. The three-dimensional power spectrum measured from the APM Galaxy Survey-2. Use of the two-dimensional power spectrum. *Mon. Not. R. Astron. Soc.* **267**, 323–332 (1994).
- Bardeen, J. M., Bond, J. R., Kaiser, N. & Szalay, A. S. The statistics of peaks of Gaussian random fields. *Astrophys. J.* **304**, 15–61 (1986).
- Abell, G. O. The distribution of rich clusters of galaxies. *Astrophys. J. Suppl.* **3**, 211–288 (1958).
- Ebeling, H. et al. Properties of the X-ray-brightest Abell-type clusters of galaxies (XBACs) from ROSAT All-Sky Survey data-I. The sample. *Mon. Not. R. Astron. Soc.* **281**, 799–829 (1996).
- Scaramella, R. Mass fluctuations on 600/h Mpc—A result from clusters of galaxies. *Astrophys. J.* **390**, L57–L60 (1992).
- Scaramella, R. et al. The ESO Slice Project [ESP] galaxy redshift survey. V. Evidence for a D=3 sample dimensionality. *Astron. Astrophys.* **334**, 404–408 (1998).
- Broadhurst, T. J., Ellis, R. S., Koo, D. C. & Szalay, A. S. Large-scale distribution of galaxies at the Galactic poles. *Nature* **343**, 726–728 (1990).
- Landy, S. D. et al. The two-dimensional power spectrum of the Las Campanas Redshift Survey: Detection of excess power on 100 h^{-1} Mpc scales. *Astrophys. J.* **456**, L1–L4 (1996).
- Kaiser, N. & Peacock, J. A. Power-spectrum analysis of one-dimensional redshift surveys. *Astrophys. J.* **379**, 482–506 (1991).
- Einasto, J. et al. A 120-Mpc periodicity in the three-dimensional distribution of galaxy superclusters. *Nature* **385**, 139–141 (1997).
- Webster, A. The clustering of radio sources. I—The theory of power-spectrum analysis. II—The 4C, GB and MC1 surveys. *Mon. Not. R. Astron. Soc.* **175**, 61–83 (1976).
- Condon, J. J. in *New Horizons from Multi-wavelength Sky Surveys* (eds McLean, B. J. et al.) 19–25 (Proc. IAU Symp. 179, Kluwer, Dordrecht, 1998).
- Shaver, P. A. & Pierre, M. Large-scale anisotropy in the sky distribution of extragalactic radio sources. *Astron. Astrophys.* **220**, 35–41 (1989).
- Benn, C. R. & Wall, J. V. Structure on the largest scales: constraints from the isotropy of radio source counts. *Mon. Not. R. Astron. Soc.* **272**, 678–698 (1995).
- Koelman, B. L., Burns, J. O. & Klypin, A. A. Two-point angular correlation function for the Green Bank 4.85 GHz Sky Survey. *Astrophys. J.* **448**, 500–509 (1995).
- Scotche, H. *Large-scale Clustering of Radio Sources from the Green Bank 1987 Survey*. Thesis, Princeton Univ. (1995).
- Loan, A. J., Wall, J. V. & Lahav, O. The correlation function of radio sources. *Mon. Not. R. Astron. Soc.* **286**, 994–1002 (1997).
- Cress, C. M., Helfand, D. J., Becker, R. H., Gregg, M. D. & White, R. L. The angular two-point correlation function for the FIRST radio survey. *Astrophys. J.* **473**, 7–14 (1996).
- Magliocchetti, M., Maddox, S. J., Lahav, O. & Wall, J. V. Variance and skewness in the FIRST survey. *Mon. Not. R. Astron. Soc.* **300**, 257–268 (1998).
- Baleisis, A., Lahav, O., Loan, A. J. & Wall, J. V. Searching for large scale structure in deep radio surveys. *Mon. Not. R. Astron. Soc.* **297**, 545–558 (1998).
- Boldt, E. A. The cosmic X-ray background. *Phys. Rep.* **146**, 215–257 (1987).
- Fabian, A. C. & Barcons, X. The origin of the X-ray background. *Annu. Rev. Astron. Astrophys.* **30**, 429–456 (1992).
- Rees, M. in *Objects of High Redshift* (eds Abell, G. O. & Peebles, P. J. E.) 207–224 (Proc. IAU Symp. 92, Reidel, Dordrecht, 1980).
- Shafer, R. A. *Spatial Fluctuations in the Diffuse Cosmic X-ray Background*. Thesis, NASA Goddard Space Flight Center (1983).
- Boughn, S. P., Crittenden, R. G. & Turok, N. G. Correlations between the cosmic X-ray and microwave backgrounds: constraints on a cosmological constant. *New Astron.* **3**, 275–291 (1998).
- Lahav, O., Piran, T. & Treyer, M. A. The X-ray background as a probe of density fluctuations at high redshift. *Mon. Not. R. Astron. Soc.* **284**, 499–506 (1997).
- Barcons, X., Fabian, A. C. & Carrera, F. J. Measuring the power spectrum of density fluctuations at intermediate redshift with X-ray background observations. *Mon. Not. R. Astron. Soc.* **293**, 60–70 (1998).
- Jahoda, K. Present epoch plus—an X-ray survey for cosmology. *Astron. Nachrichten* **319**, 129–132 (1998).
- Boughn, S. P. Cross-correlation of the 2–10 keV X-ray background with radio sources: Constraining the large-scale structure of the X-ray background. *Astrophys. J.* **499**, 533–541 (1998).
- Treyer, M. et al. Large scale fluctuations in the X-Ray Background. *Astrophys. J.* **509**, 531–536 (1998).
- Fry, J. N. The evolution of bias. *Astrophys. J.* **461**, L65–L68 (1996).
- Shanks, T. & Boyle, B. J. QSO clustering—I. Optical surveys in the redshift range $0.3 < z < 2.2$. *Mon. Not. R. Astron. Soc.* **271**, 753–772 (1994).
- Steidel, C. C. et al. A large structure of galaxies at redshift z approximately 3 and its cosmological implications. *Astrophys. J.* **492**, 428–438 (1998).
- Webb, J. K. & Barcons, X. A search for inhomogeneities in the Lyman-alpha forest. *Mon. Not. R. Astron. Soc.* **250**, 270–277 (1991).
- Villumsen, J. V. Weak lensing by large-scale structure in open, flat and closed universes. *Mon. Not. R. Astron. Soc.* **281**, 369–383 (1996).
- Kaiser, N. Weak lensing and cosmology. *Astrophys. J.* **498**, 26–42 (1998).
- Strauss, M. A., Yahil, A., Davis, M., Huchra, J. P. & Fisher, K. A redshift survey of IRAS galaxies. V—The acceleration on the Local Group. *Astrophys. J.* **397**, 395–419 (1992).
- Webster, M., Lahav, O. & Fisher, K. Wiener reconstruction of the IRAS 1.2-Jy galaxy redshift survey: cosmographical implications. *Mon. Not. R. Astron. Soc.* **287**, 425–444 (1997).
- Riess, A. G., Davis, M., Baker, J. & Kirshner, R. P. The velocity field from type Ia supernovae matches the gravity field from galaxy surveys. *Astrophys. J.* **488**, L1–L6 (1997).
- Sachs, R. K. & Wolfe, A. M. Perturbations of a cosmological model and angular variations of the microwave background. *Astrophys. J.* **147**, 73–90 (1967).
- White, M., Scott, D. & Silk, J. Anisotropies in the cosmic microwave background. *Annu. Rev. Astron. Astrophys.* **32**, 319–370 (1994).
- Scott, D., Silk, J. & White, M. From microwave anisotropies to cosmology. *Science* **268**, 829–835 (1995).
- Hu, W., Sugiyama, N. & Silk, J. The physics of microwave background anisotropies. *Nature* **386**, 37–43 (1997).
- Bennett, C. L., Turner, M. S. & White, M. The cosmic rosetta stone. *Phys. Today* **50**, 32–38 (1997).
- Gawiser, E. & Silk, J. Extracting primordial density fluctuations. *Science* **280**, 1405–1411 (1998).
- Webster, M. et al. Joint estimation of cosmological parameters from CMB and IRAS data. *Astrophys. J.* **509**, L65–L68 (1998).
- Stoeger, W. R., Maartens, R. & Ellis, G. F. R. Proving almost-homogeneity of the universe: An almost Ehlers-Green-Sachs theorem. *Astrophys. J.* **443**, 1–5 (1995).
- Seljak, U. Gravitational lensing effect on cosmic microwave background anisotropies: A power spectrum approach. *Astrophys. J.* **463**, 1–7 (1996).
- Lynden-Bell, D. et al. Spectroscopy and photometry of elliptical galaxies. V—galaxy streaming toward the new supergalactic center. *Astrophys. J.* **326**, 19–49 (1988).
- Dekel, A. Dynamics of cosmic flows. *Annu. Rev. Astron. Astrophys.* **32**, 371–418 (1994).
- Strauss, M. A. & Willick, J. A. The density and peculiar velocity fields of nearby galaxies. *Phys. Rep.* **261**, 271–431 (1995).
- Kolatt, T. & Dekel, A. Large-scale power spectrum from peculiar velocities. *Astrophys. J.* **479**, 592–605 (1997).
- Lauer, T. R. & Postman, M. The motion of the Local Group with respect to the 15000 kilometer per second Abell cluster inertial frame. *Astrophys. J.* **425**, 418–438 (1994).
- Guzzo, L., Iovino, A., Chincarini, G., Giovanelli, R. & Haynes, M. P. Scale-invariant clustering in the large-scale distribution of galaxies. *Astrophys. J.* **382**, L5–L9 (1991).
- Martinez, V. J. & Coles, P. Correlations and scaling in the QDOT redshift survey. *Astrophys. J.* **437**, 550–555 (1994).
- Lemson, G. & Sanders, R. H. On the use of the conditional density as a description of galaxy clustering. *Mon. Not. R. Astron. Soc.* **252**, 319–328 (1991).
- Cappi, A., Benoist, C., Da Costa, L. N. & Maurogordato, S. Is the Universe a fractal? Results from the Southern Sky Redshift Survey 2. *Astron. Astrophys.* **335**, 779–788 (1998).
- Martinez, V. J., Pons-Borderia, M.-J., Moyeed, R. A. & Graham, M. J. Searching for the scale of homogeneity. *Mon. Not. R. Astron. Soc.* **298**, 1212–1222 (1998).
- Guzzo, L. Is the universe homogeneous? (On large scales). *New Astron.* **2**, 517–532 (1997).
- Calzetti, D., Giallisco, M., Ruffini, R., Taraglio, S. & Bahcall, N. A. Clustering of galaxies: Fractal or homogeneous infrastructure? *Astron. Astrophys.* **245**, 1–6 (1991).
- Maartens, R., Ellis, G. F. R. & Stoeger, W. R. Anisotropy and inhomogeneity of the universe from $\Delta T/T$. *Astron. Astrophys.* **309**, L7–L10 (1996).
- Ellis, G. F. R. Alternatives to the Big Bang. *Annu. Rev. Astron. Astrophys.* **22**, 157–184 (1984).
- Saunders, W., Rowan-Robinson, M. & Lawrence, A. The spatial correlation function of IRAS galaxies on small and intermediate scales. *Mon. Not. R. Astron. Soc.* **258**, 134–146 (1992).
- Bahcall, N. A. Large-scale structure in the universe indicated by galaxy clusters. *Annu. Rev. Astron. Astrophys.* **26**, 631–686 (1988).
- Kaiser, N. On the spatial structures of Abell clusters. *Astrophys. J.* **284**, L9–L12 (1984).
- Dressler, A. Galaxy morphology in rich clusters—Implications for the formation and evolution of galaxies. *Astrophys. J.* **236**, 351–365 (1980).
- Loveday, J., Maddox, S. J., Efstathiou, G. & Peterson, B. A. The Stromlo-APM redshift survey. 2: Variation of galaxy clustering with morphology and luminosity. *Astrophys. J.* **442**, 457–468 (1995).
- Hermit, S. et al. The two-point correlation function and morphological segregation in the Optical Redshift Survey. *Mon. Not. R. Astron. Soc.* **283**, 709–720 (1996).
- Dekel, A. & Rees, M. J. Physical mechanisms for biased galaxy formation. *Nature* **326**, 455–462 (1987).
- Babul, A. & White, S. D. M. Quasar-modulated galaxy clustering in a cold dark matter universe. *Mon. Not. R. Astron. Soc.* **253**, 31P–34P (1991).
- Coles, P. Galaxy formation with a local bias. *Mon. Not. R. Astron. Soc.* **262**, 1065–1075 (1993).
- Bower, R. G., Coles, P., Frenk, C. S. & White, S. D. M. Cooperative galaxy formation and large-scale structure. *Astrophys. J.* **405**, 403–412 (1993).
- Kauffmann, G., Nusser, A. & Steinmetz, M. Galaxy formation and large-scale bias. *Mon. Not. R. Astron. Soc.* **286**, 795–811 (1997).
- Bagla, J. Evolution of galaxy clustering. *Mon. Not. R. Astron. Soc.* **299**, 417–424 (1998).
- Dekel, A. & Lahav, O. Stochastic nonlinear galaxy biasing. *Astrophys. J.* (in the press); also as preprint astro-ph/9806193 at (<http://xxx.lanl.gov>) (1998).
- Szalay, A. S. & Schramm, D. N. Are galaxies are more strongly correlated than clusters? *Nature* **314**, 718–719 (1985).
- Martinez, V. J., Jones, B. J. T., Dominguez-Tenreiro, R. & Van De Weygaert, R. Clustering paradigms and multifractal measures. *Astrophys. J.* **357**, 50–61 (1990).
- Smoot, G. F. et al. Structure in the COBE differential microwave radiometer first-year maps. *Astrophys. J.* **396**, L1–L5 (1992).
- Bennett, C. L. et al. Four-year COBE DMR cosmic microwave background observations: maps and basic results. *Astrophys. J.* **464**, L1–L4 (1996).
- Hancock, S. et al. Studies of cosmic microwave background structure at Dec.=+40 deg - II. Analysis and cosmological interpretation. *Mon. Not. R. Astron. Soc.* **289**, 505–514 (1997).

Acknowledgements. We thank A. Baleisis, E. Gawiser, I. Guzzo and M. Treyer for discussions.

Correspondence and requests for materials should be addressed to K.K.S.W. (e-mail: kwu@ast.cam.ac.uk).

The generation of martian floods by the melting of ground ice above dykes

Dan McKenzie & Francis Nimmo

Institute of Theoretical Geophysics, Bullard Laboratories, Department of Earth Sciences, University of Cambridge, Madingley Road, Cambridge CB3 0EZ, UK

The surface of Mars is cut by long linear faults with displacements of metres to kilometres, most of which are thought to have been formed by extension¹. The surface has also been modified by enormous floods, probably of water, which often flowed out of valleys formed by the largest of these faults². By analogy with structures on Earth^{3,4}, we propose here that the faults are in fact the surface expression of dykes⁵, and not of large-scale tectonic movements. We use a numerical model to show that the intrusion of large dykes can generate structures like Valles Marineris. Such dykes can provide a heat source to melt ground ice, and so provide a source of water for the floods that have been inferred to originate in some of the large valleys².

Most silica-rich bodies in the Solar System have reached their melting temperature at some stage in their history⁶, and the resulting surface features, such as volcanoes, lava flows and lava channels, have been mapped in some detail on Venus, Mercury, the Moon and Mars. There has been less interest in subsurface volcanic features, such as dykes⁷. On Earth, intrusive and extrusive volcanism generally occur together, and the same is likely to be true on other planets. The linearity and enormous horizontal extent of dykes on Earth is quite unlike structures produced by tectonic deformation. Terrestrial mafic dykes³ are up to 2,000 km long, with widths of 50–100 m and volumes of $\sim 500 \text{ km}^3$. They must be intruded in a few tens of days if the magma is not to solidify. When such an event occurs, a small collapse graben forms if the pressure of the magma is not sufficient for it to reach the surface. Depressions produced by the intrusion of small dykes have been mapped in Iceland⁸, where their surface expression is dominated by the presence of larger faults resulting from plate separation. On Venus, depressions associated with regional dykes produce surface roughness that is clearly visible on radar images^{7,9}, and can be followed for thousands of kilometres. Mège and Masson⁵ have proposed that similar features on Mars result from dyke intrusion. Though such graben have generally been mapped as tectonic features¹, their displacement–length ratios are much smaller than those of normal faults associated with tectonic extension¹⁰. When the same scales are used, maps of graben swarms

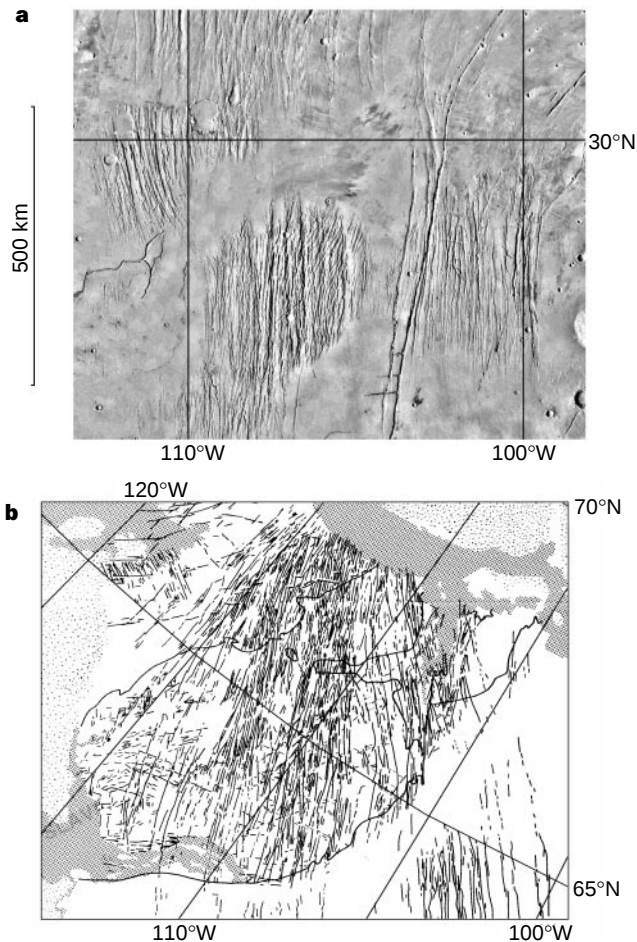


Figure 1 A comparison of graben swarms on Mars with dyke swarms on Earth. **a**, A mosaic of the northeastern portion of the Tharsis quadrangle (see Fig. 2a), showing graben swarms from several different generations of faulting (modified from USGS data available at <http://www.pdsimage.wr.usgs.gov/PDS/public/mapmaker/mapmkr.htm>). **b**, Part of the Mackenzie dyke swarm in northwest Canada (see Fig. 2b), based on ref. 4 and shown at the same scale as **a**. The boundaries of several metamorphic provinces of the Canadian Shield are also shown as wavy lines.

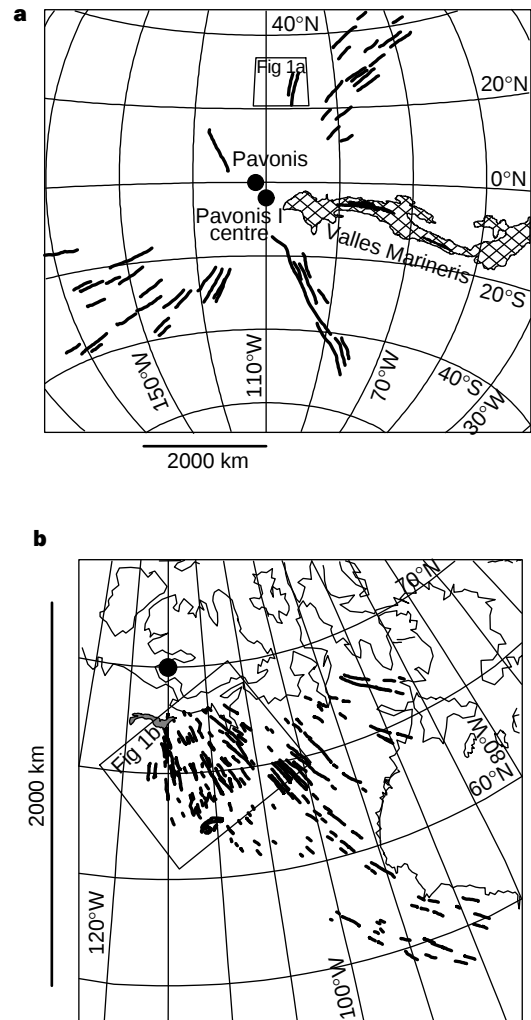


Figure 2 A comparison of two regions of Mars and Earth, both dominated by convective plumes. The large filled circles mark the likely positions of the plume centres that formed the radial structures. **a**, Azimuthal equidistant projection on 4°S, 110°W of fractures associated with the Pavonis I centre¹. **b**, Mackenzie dyke swarm⁴ using the same projection with centre 70°N, 115°W.

on Mars (Fig. 1a) closely resemble dyke swarms on Earth (Fig. 1b), though the longest graben on Mars (Fig. 2a) are about twice the length of the longest terrestrial dykes (Fig. 2b). On Earth, large dyke swarms—such as the Mackenzie swarm³ shown in Figs 1b and 2b—originate from the sites of plumes, where melt is produced by localized upwelling of hot mantle¹¹. On Mars, the large volcano Pavonis Mons, from which the graben in Fig. 2a radiate, has been constructed on top of the Tharsis bulge, which has the topographic and gravity anomalies expected from a large plume¹².

Plescia and Saunders¹ noticed that the Valles Marineris is radial to the Pavonis 1 centre in Fig. 2a. Therefore this large graben system may also be the surface expression of an enormous dyke. The cross-sectional area of the dyke modelled below is 350 km², and hence has a volume of 10⁶ km³ if its length is 3,000 km. The volume of one of the largest dykes on Earth, the Great Dyke of Zimbabwe¹³, is not likely to be greater than 5×10^4 km³, or a factor of 20 smaller than that proposed for Mars. However, the volume of Olympus Mons on Mars is about a factor of 40 greater than that of Hawaii, the largest volcano on Earth, and the volume of melt produced by terrestrial plumes¹¹ is generally between 10⁶ and 10⁷ km³. The melt volume required by the model in Fig. 3 is therefore not unreasonably large. Like the features on Mars, the Great Dyke is associated with a number of parallel smaller dykes¹³. A large dyke could not only account for the linearity of Valles Marineris, but could also provide an explanation for the melting of ground ice which has produced the huge volumes of water that intermittently flooded out from the eastern end of this graben system to collect in the low-lying Chryse Planitia². Similar floods have flowed out of other valleys on Mars, such as Mangala Valles¹⁴, which also have graben at their probable

sources. Baker² has compared the features produced by these floods with the scablands of Washington State, USA, and has argued that the flow rates on Mars must have been $\sim 2 \times 10^7$ m³ s⁻¹, with a total volume of water of $\sim 2,000$ km³. In many places the margins of Valles Marineris consist of curved scarps. This morphology has been produced by mass wasting, which has caused the scarps to retreat from the large straight faults that bound the deeper parts of the troughs¹⁵. Masursky *et al.*¹⁶ argued that the scarp retreat was caused by the melting of a layer of ground ice, producing springs at the base of the escarpment that remove material and so cause landslides. They also suggested that the melting was caused by volcanism.

The processes involved in the transfer of volcanic heat to permafrost have received little attention¹⁷. The cooling of 1 kg of basalt from a temperature of 1,500 to 200 K releases enough heat to melt 5 kg of ice. Therefore a dyke 2,000 km long, 5 km deep and 50 m across, similar to one of the Mackenzie dykes, contains enough heat to produce 7,500 km³ of water. Hence there is no obvious energetic problem in generating the required volumes of water by dyke intrusion. We used a simple finite-difference model to discover whether the proposed mechanism could account for the general behaviour observed, rather than to model any individual feature. Figure 3 shows that intrusion of a dyke with a width of 16 km and a depth extent of 22 km, whose top is 9 km below the surface, produces a graben 26 km across and 6 km deep if the bounding faults dip at an angle of 60°. The base of the permafrost layer (5 km thick with an ice fraction of 0.2 by volume) first melts above the dyke. The melting zone spreads vertically and laterally with time, but never reaches the top of the frozen layer. At its greatest horizontal extent, the melt water extends over a width that is about twice that of the graben. Because the temperature is buffered at the melting point of ice until all the ice has melted, the surface temperature gradient remains low after dyke intrusion. Hence the melt water at the base of the permafrost does not start to refreeze until 8 Myr after the dyke is intruded. The vertical displacement on the faults which bound the graben is greater than the depth to the base of the permafrost, and melt water draining into the graben will lead to scarp retreat. In the absence of such draining, the pore pressure in the water will equal the weight of the overburden, because upward percolation is prevented by the permafrost. Regions underlain by such overpressured water are liable to large-scale catastrophic failure that produces a chaotic terrain like that commonly found on Mars¹⁵. The total volume of water generated reaches 6.5 km³ for each 1-km length of dyke 8 Myr after dyke intrusion. A 2,000-km-long dyke could therefore generate 1.3×10^4 km³ of water. When the fraction of ice is 0.2, the subsidence caused by the melting is small compared to that produced by the dyke intrusion. Sills are also likely to occur on Mars, but would not produce any obvious surface expression. However, they would be at least as effective as dykes in melting ground ice.

It is difficult to model melting on Mars in detail, because the relevant parameters are poorly known. Estimates of the age of the floods², of 2–3 Gyr, are also uncertain, and depend on calibration of the cratering rate. The proportion of ice is unknown, though various estimates have been made¹⁸. The fraction could be as high as 1 in some depth intervals if frozen lakes are present. The thickness of the permafrost layer is controlled by the surface temperature gradient, and hence by the heat flux. Measurements of ¹⁴²Nd/¹⁴⁴Nd, ¹⁴³Nd/¹⁴⁴Nd, ¹⁷⁶Hf/¹⁷⁷Hf and ⁸⁷Sr/⁸⁶Sr in SNC (Shergotty–Nakhla–Chassigny) meteorites^{19,20} require the martian mantle to have been depleted in incompatible elements by melting within a few hundreds of Myr of its formation. This process transported almost all of the initial inventory of the heat-producing elements (K, U and Th) into the martian crust, where the isotopic constraints require them to have remained thereafter. Convective models of martian thermal evolution that are compatible with the geochemical observations suggest that heat flow from the mantle 2 Gyr ago was only

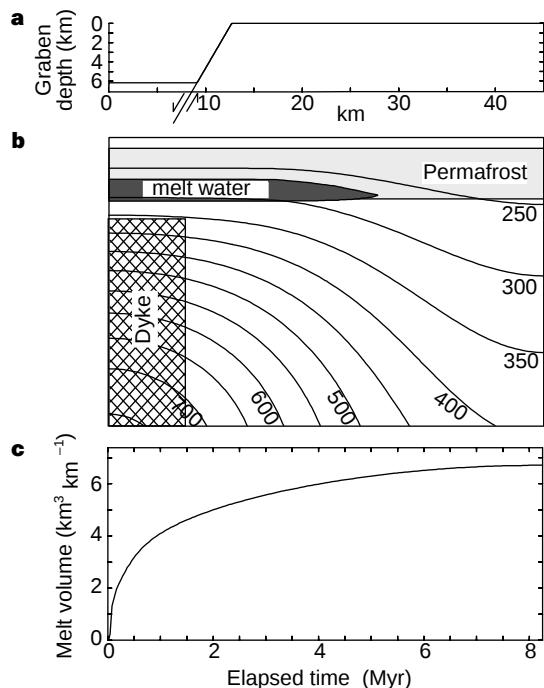


Figure 3 A numerical model of melting of ground ice by dyke intrusion. The calculation was performed using a rectangular box on a square mesh of 60×40 , plotted with the same vertical and horizontal scales. The left- and right-hand sides of the box are mirror planes, the heat flux through the base is fixed at 16 mW m^{-2} , and the surface temperature is fixed at 200 K. The conductivity is $2 \text{ W m}^{-1} \text{ K}^{-1}$, the initial temperature of the dyke is 1,500 K, the latent heat of basalt is 560 kJ kg^{-1} , the specific heat is $1.2 \text{ kJ kg}^{-1} \text{ K}^{-1}$, and the fraction of ice present in the permafrost layer is 0.2. **a**, Cross-section of the surface graben, calculated by area conservation with fault dips of 60°. The faults start at the top corners of the dyke. Subsidence from melting is small and is neglected. **b**, Isotherms in K. **c**, Melt water volume per kilometre of dyke. After 8.25 Myr, the water starts to refreeze.

$\sim 16 \text{ mW m}^{-2}$, or less than half of that expected in the absence of geochemical differentiation¹². The surface temperature gradient depends on the integrated crustal radioactivity as well as the mantle heat flux, and may vary from 6 K km^{-1} (where the crustal contribution is not important) to more than 20 K km^{-1} (where it dominates). If the surface temperature is 200 K , the base of the permafrost is at a depth of between 11 and 3 km . Low geothermal gradients early in martian history are also suggested by geophysical estimates of the effective elastic thickness of more than 100 km beneath the Tharsis volcanoes²¹.

The large volumes of flood water that Baker² and Marsursky *et al.*¹⁶ propose to account for the observed landforms must have been released suddenly. The melting process modelled in Fig. 3 generates water far too slowly to maintain flow rates of $10^7 \text{ m}^3 \text{ s}^{-1}$. The melt water must therefore first collect, either above or below the surface, before being suddenly released. Valles Marineris contains a number of closed depressions, some of which are dammed by large landslides. Such structures could collect water released from the subsurface, and are likely to undergo catastrophic failure, as they do on Earth.

The mass of water generated by the model in Fig. 3 is comparable to the total mass of the martian atmosphere. Spreading such a layer of water over a large part of Chryse Planitia might therefore have affected the surface temperature and humidity of the entire planet²². Most craters in Chryse Planitia show evidence of remobilization of ground ice²³. This discussion shows that there are no obvious problems with the proposed dyke model, which can account for graben such as Valles Marineris and which can produce the necessary volumes of water for the martian floods. If the above arguments are correct, only a few relatively minor structures on Mars, such as Thaumasia graben¹, have a purely tectonic origin. □

Received 28 January; accepted 19 October 1998.

1. Plescia, J. B. & Saunders, R. S. Tectonic history of the Tharsis Region, Mars. *J. Geophys. Res.* **87**, 9775–9791 (1982).
2. Baker, V. R. *The Channels of Mars* (Univ. Texas Press, Austin, 1982).
3. Fahrig, W. F. in *Mafic Dyke Swarms* (eds Halls, H. C. & Fahrig, W. F.) 331–348 (Spec. Pap. 34, Geol. Assoc. Canada, St John's, Newfoundland, 1987).
4. Fahrig, W. F. & West, T. D. Diabase dyke swarms of the Canadian Shield. (Map 1627A, Geol. Surv. Canada, Ottawa, 1986).
5. Mège, D. & Masson, P. A plume tectonics model for the Tharsis province, Mars. *Planet. Space Sci.* **44**, 1499–1546 (1996).
6. Basaltic Volcanism Study Project *Basaltic Volcanism on Terrestrial Planets* (Pergamon, New York, 1983).
7. Ernst, R. E., Head, J. W., Parfitt, E. A., Grosfils, E. & Wilson, L. Giant radiating dyke swarms on Earth and Venus. *Earth Sci. Rev.* **39**, 1–58 (1995).
8. Sigurdsson, O. Surface deformation of the Krafla fissure swarm in two rifting events. *J. Geophys. Res.* **47**, 154–159 (1980).
9. McKenzie, D., McKenzie, J. M. & Saunders, R. S. Dike emplacement on Venus and on Earth. *J. Geophys. Res.* **97**, 15977–15990 (1992).
10. Schultz, R. A. Displacement-length scaling for terrestrial and Martian faults: Implications for Valles Marineris and shallow planetary grabens. *J. Geophys. Res.* **102**, 12009–12015 (1997).
11. White, R. S. & McKenzie, D. Mantle plumes and flood basalts. *J. Geophys. Res.* **100**, 17543–17585 (1995).
12. Schubert, G., Solomon, S. C., Turcotte, D. L., Drake, M. J. & Sleep, N. H. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 147–183 (Univ. Arizona Press, Tucson, 1992).
13. Worst, B. G. The Great Dyke of Southern Rhodesia. *S. Rhodesia Geol. Surv. Bull.* **47**, (1960).
14. Zimbelman, J. R., Craddock, R. A., Greeley, R. & Kuzmin, R. O. Volatile history of Mangala Valles, Mars. *J. Geophys. Res.* **97**, 18309–18317 (1992).
15. Lucchitta, B. K. *et al.* in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 453–492 (Univ. Arizona Press, Tucson, 1992).
16. Masursky, H., Boyce, J. M., Dial, A. L., Schaber, G. G. & Strobell, M. E. Classification and time of formation of Martian channels based on Viking data. *J. Geophys. Res.* **82**, 4016–4038 (1977).
17. Squyres, S. W., Wilhelms, D. E. & Moosman, A. C. Large-scale volcano-ground ice interaction on Mars. *Icarus* **70**, 385–408 (1987).
18. Squyres, S. W., Clifford, S. M., Kuzmin, R. O., Zimbelman, J. R. & Costard, F. M. in *Mars* (eds Kieffer, H. H., Jakosky, B. M., Snyder, C. W. & Matthews, M. S.) 523–556 (Univ. Arizona Press, Tucson, 1992).
19. Borg, L. E., Nyquist, L. E., Taylor, L. A., Wiesmann, H. & Shih, C.-Y. Constraints on Martian differentiation processes from Rb–Sr & Sm–Nd isotopic analyses of the basaltic Shergottite QUE94201. *Geochim. Cosmochim. Acta* **61**, 4915–4931 (1997).
20. Blichert-Toft, J. *et al.* Lu–Hf isotopic compositions of SNC meteorites: Implications of Martian mantle evolution. *Mineral. Mag.* **62**, 168–169 (1998).
21. Solomon, S. C. & Head, J. W. Heterogeneities in the thickness of the elastic lithosphere on Mars: Constraints on heat flow and internal dynamics. *J. Geophys. Res.* **95**, 11073–11083 (1990).
22. Baker, V. R. *et al.* Ancient oceans, ice sheets and the hydrological cycle on Mars. *Nature* **352**, 589–594 (1991).
23. Carr, M. H. *et al.* Martian impact craters and emplacement of ejecta by surface flow. *J. Geophys. Res.* **82**, 4055–4065 (1977).

Acknowledgements. We thank J. McKenzie for his help, and the Royal Society, Magdalene College, Cambridge, and NERC for support.

Correspondence and requests for materials should be addressed to F.N. (e-mail: nimmo@esc.cam.ac.uk).

Controlling the shape of a quantum wavefunction

T. C. Weinacht, J. Ahn & P. H. Bucksbaum

Department of Physics, University of Michigan, Ann Arbor, Michigan 48109-1120, USA

The ability to control the shape and motion of quantum states^{1,2} may lead to methods for bond-selective chemistry and novel quantum technologies, such as quantum computing. The classical coherence of laser light has been used to guide quantum systems into desired target states through interfering pathways^{3–5}. These experiments used the control of target properties—such as fluorescence from a dye solution⁶, the current in a semiconductor^{7,8} or the dissociation fraction of an excited molecule⁹—to infer control over the quantum state. Here we report a direct approach to coherent quantum control that allows us to actively manipulate the shape of an atomic electron's radial wavefunction. We use a computer-controlled laser to excite a coherent state in atomic caesium. The shape of the wavefunction is then measured¹⁰ and the information fed back into the laser control system, which reprograms the optical field. The process is iterated until the measured shape of the wavefunction matches that of a target wavepacket, established at the start of the experiment. We find that, using a variation of quantum holography¹¹ to reconstruct the measured wavefunction, the quantum state can be reshaped to match the target within two iterations of the feedback loop.

As in classical holography, quantum holography relies on interference between two waves, which can be labelled as 'object' and 'reference'. Whereas classical holography relies on classical light waves diffracting from scattering media, in quantum holography the object and reference are combined in a single coherently prepared quantum state. This state is a superposition of a shaped wavepacket (the 'object') and a reference wavepacket (the 'signal'). The object and reference packets are created by two specially shaped coherent laser pulses with a time delay, τ , between them. The interference between the object and the reference modulates the spectrum of the wavefunction. A careful measurement of this interference and how it changes with the time delay τ can be used to reconstruct the object wavefunction¹⁰. This is similar to the optical technique known as spectral interferometry, which is used to determine the phase and amplitude of a classical light pulse^{12,13}.

The essential feature in our application of quantum holography is the ability to measure independently the amplitude of each state in the total wavepacket. This is accomplished with state selective field ionization (SSFI)¹⁴. SSFI allows us to project the total wavefunction onto the eigenstates of the atomic hamiltonian and directly measure the amplitude of each projection. Eigenstate signals indicate the amplitude of each state in the wavepacket. The correlations in fluctuations of pairs of states indicate the relative phase difference between two states in the object wavepacket and their corresponding states in the reference packet^{15,16}. The measured phase and amplitude of each state can be used to reconstruct the object wavepacket as a coherent superposition of eigenstates^{17–20}. In the work reported here, this measurement was used as feedback to reshape the optical excitation pulse and thereby reshape the wavepacket.

Both the signal and reference wavepackets were created by irradiating an effusive beam of prepared caesium atoms with two ultrafast, broadband light pulses from an amplified titanium sapphire laser system. The atoms were prepared in the $7s$ state from the $6s$ ground state via a two-photon transition driven by an intense nanosecond dye-laser pulse with a wavelength of $1.08 \mu\text{m}$. The ultrafast pulses excited a small portion of the electronic wavefunc-

tion (consistent with first-order perturbation theory) to the np Rydberg series. Both ultrafast pulses originated from the same Kerr-lens mode-locked (KLM) titanium sapphire laser. The output of the KLM laser was sent through a dispersive grating-pulse expander²¹ and amplified in a 10-Hz regenerative amplifier. The light from the amplifier was split to form two pulses. One of these pulses was directed into an optical pulse shaper, which consisted of a spectrally resolved Bragg deflector at the Fourier plane of a zero-dispersion pulse stretcher²². The acoustic waveform on the Bragg deflector was programmable, with independent control of phase and amplitude. This allowed us to reshape the optical pulse to any field consistent with the available optical and acoustic bandwidth²². To compensate for the inefficiency of the pulse shaper (<10%), the shaped pulse was reamplified in a low-gain, linear multipass amplifier and then recompressed in a parallel grating compressor. The second pulse from the regenerative amplifier was not shaped or reamplified but simply recompressed in a second parallel grating compressor to yield a 150-fs nearly transform-limited pulse. The two pulses were recombined on an optical beam splitter and sent into the vacuum chamber containing the atomic beam.

Our measurements were possible despite the fact that the shaped and reference pulses were not phase-locked. The absence of phase-locking meant that the first-order coherence in the field formed by the two pulses averaged to zero after multiple laser shots. We therefore looked at the second-order coherence of the wavepacket consisting of correlations in the fluctuations of the eigenstate populations resulting from interference between the two excitations. This is essentially a Hanbury-Brown and Twiss measurement applied to atoms¹⁶. These authors used correlations in the temporal fluctuations of starlight observed by two spatially separate telescopes to measure the size of a star. In the case of atoms, we used spectrally resolved fluctuations of electronic probability to measure the phase of the electronic wavefunction.

The total wavepacket created by the coherent sum of the signal

and reference for each laser shot can be written as:

$$\Psi_{\text{total}}(t = \tau) = \sum_{n=24}^{31} (a_n e^{-i\omega_n \tau} + b_n e^{-i\omega_{\text{gs}} \tau}) \psi_n$$

where $a_n = |a_n| e^{i\delta_n}$. The complex coefficients a_n were determined by the shaped laser pulse and the real b_n s were determined by the spectrum of the transform-limited reference pulse. The ψ_n represent the atomic Rydberg np eigenstates with principal quantum number n ranging from 24 to 31. The phase acquired by each state during the time delay τ between the two pulses ($\omega_n \tau$) depended on the energy of each state. The phase acquired by the ground state during this same time delay was ($\omega_{\text{gs}} \tau$).

The population of each eigenstate in the wavepacket (the square modulus of the projection of the total wavefunction onto the eigenstate) was measured by SSFI. This consisted of exposing the Rydberg atoms in the beam to a linearly increasing electric field as a function of time. Each eigenstate in the wavepacket had a different field ionization threshold, and hence ionized at a different time. The ionization signal as a function of time, as measured by an electron multiplier, displayed a series of peaks whose integrals were proportional to the population of each eigenstate. The population of each eigenstate as measured by SSFI was:

$$P_n = |a_n|^2 + |b_n|^2 + 2|a_n||b_n|\cos((\omega_n - \omega_{\text{gs}})\tau - \delta_n)$$

In the absence of a reference pulse, the coefficients P_n still yield the $|a_n|$ directly, but without the second wavepacket the P_n cannot give any information about the phase of these complex coefficients.

The destruction of first-order coherence over several laser pulses is due to small random variations in τ of the order of an optical period. These fluctuations can be caused by slight air currents and table vibrations. In our experiment, there was also a random phase introduced between signal and reference pulses by the carrier

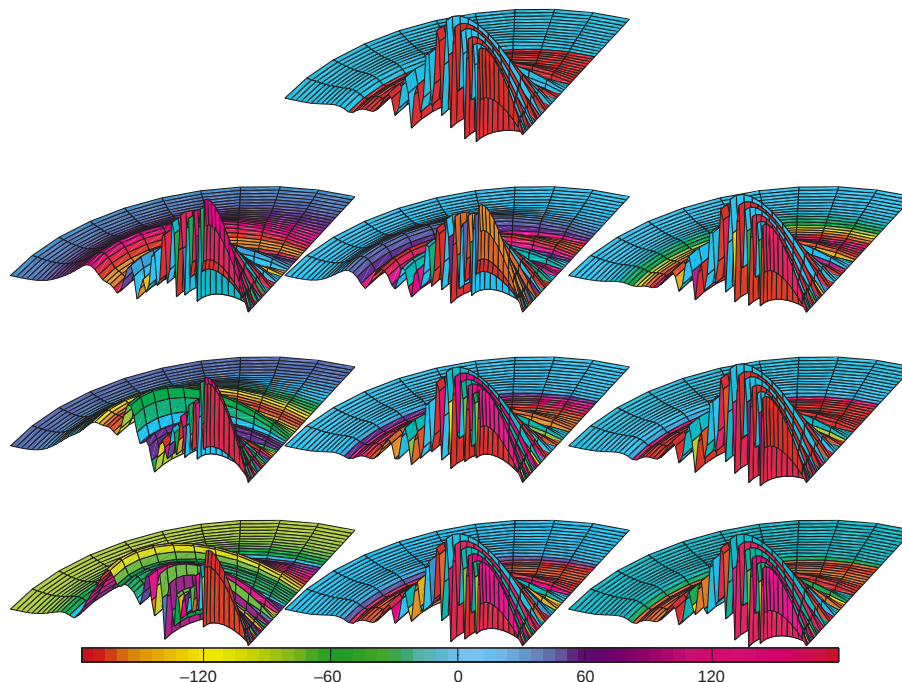


Figure 1 Atomic Rydberg wavefunctions reconstructed from measured projections. The top panel shows the target wavefunction which is the only wavefunction that was not measured. For the other panels, new rows show wavefunctions measured at new time delays, and columns show successive iterations in the feedback loop. The axes are not shown explicitly in order to conserve space but they are as follows: the vertical axis shows the amplitude of

the wavefunction in atomic units (1 a.u. = 0.529 Å), colour denotes phase, the axis to the right is the z axis and the other is the x axis. The range of the x and z axes is 1,800 a.u. and the vertical axis range is 1.5×10^{-3} a.u. The phase shown ranges from -180 degrees to 180 degrees and is illustrated by the colour bars shown at the bottom of the figure (in degrees). The excitation laser was polarized along the x -axis.

frequency of the acoustic wave in the Bragg deflector, which was not phase-locked to the output of the KLM laser. Over several laser shots, the cosine term in the P_n averages to zero and any interference yielding the relative phase between states δ_{nm} washes out. In general, measurements of first-order coherence oscillate at the laser frequency $\omega_L = \omega_n - \omega_{gs}$ and therefore vanish in ensemble averages with variations of delay on the order of $2\pi/\omega_L$:

$$\langle\langle\Psi_{\text{signal}}|\Psi_{\text{reference}}\rangle\rangle \propto \left\langle \sum_{n=24}^{31} \cos((\omega_n - \omega_{gs})\tau - \delta_n) \right\rangle = 0$$

Measurements involving second-order coherence contain oscillations at the Kepler frequency (frequency differences in the laser bandwidth), and therefore retain the desired phase information while averaging over several laser shots for these same temporal variations:

$$\langle\langle\Psi_{\text{signal}}|\Psi_{\text{reference}}\rangle\rangle \langle\langle\Psi_{\text{signal}}|\Psi_{\text{reference}}\rangle\rangle \propto \left\langle \sum_{n,m} \cos((\omega_n - \omega_m)\tau - \delta_{nm}) \right\rangle \neq 0$$

We made use of the second-order coherence of the wavepackets by constructing the normalized correlation coefficient between pairs of eigenstate populations:

$$r_{nm} = \frac{\langle P_n P_m \rangle - \langle P_n \rangle \langle P_m \rangle}{(\Delta P_n)(\Delta P_m)} = \cos((\omega_n - \omega_m)\tau - \delta_{nm})$$

In this expression, the factors ΔP_n are the standard deviations in the populations P_n .

This result also permitted us to measure the degree of coherence with the following procedure: we programmed the Bragg deflector to step each state in the wavepacket through a series of phases with respect to each of the other states in the wavepacket. The phase difference between any two states in the wavepacket was determined by fitting r_{nm} to a cosine curve. Decoherence of the shaped wavepacket caused a decrease in the amplitude of the measured cosine curve. Using this method, we were able to detect coherence in the shaped wavepacket lasting beyond 1 ns for atoms at a temperature of 400 K.

Knowing the amplitude and relative phase of each state in the wavepacket, the electronic wavefunction could be completely reconstructed¹⁰. Once reconstructed, the difference between the measured wavefunction and the desired target state was computed and used to reprogram the phase of each state via the Bragg deflector. Having the full reconstruction information obviated the need for a sophisticated algorithm to implement feedback, and resulted in very rapid convergence to any target that we chose. In our simple feedback algorithm, the relative phase of each projection was adjusted by an amount equal to the difference between the measured phase and the target phase. For any given delay τ , the algorithm was able to control the shaped wavefunction to match the target within two iterations of the feedback loop. Nonlinearities in the Bragg deflector response and other technical control problems were automatically corrected in the feedback loop.

The system was capable of adapting to changing initial conditions and maintaining a fixed output (the target wavefunction) with a changing input. This was demonstrated by changing the measurement time delay τ after the algorithm converged. The algorithm adapted to the new launch time and compensated for the extra dispersion introduced by the new delay such that the wavepacket matched the target at the new measurement time. The measurement delay was altered several times, and each time the algorithm converged quickly. Figure 1 shows the shaped wavepacket following each iteration. The top frame shows the target wavefunction. All other frames show measured wavefunctions at the time the reference pulse was incident on the atoms. Each row starts with a different delay setting, and each new column shows the next iteration in the feedback loop.

The rapid convergence of this experiment is due partly to the fact that the excitation of the shaped wave packet was linear in light fluence, that is, the excitation was in the weak-field limit. This is not a requirement for the measurement technique, nor is it a limitation on the application of wavefunction reconstruction to coherent control. As long as the reference pulse excites a wavepacket in the weak field limit, the phase and amplitude of any wavepacket interfering with the reference can be measured. We are currently using this technique to measure wavepackets created by half-cycle pulses²³, which interact with the atom in a highly non-perturbative fashion.

If the initial wavepacket is created in the strong-field regime, then convergence will not be as rapid as in this experiment and a more sophisticated algorithm will be required to implement control with feedback. However, knowledge of the full wavefunction can be used to direct the search for the optimal pulse shape and minimize the amount of phase space that has to be explored. Furthermore, the reconstruction provides a direct comparison between the results of the optimal pulse shape and the target.

We have demonstrated active control of a Rydberg atomic electron via holographic reconstruction. The reconstructed wavefunction guided control of its shape. Direct applications of this technique include using shaped Rydberg atoms as transient nanostructures for time-gated diffraction of X-rays²⁴ and as storage and computation modules for quantum information processing and computing. Along more general lines, this experiment serves as an illustration of the utility of quantum state reconstruction as another tool for achieving coherent control via feedback. □

Received 30 July; accepted 20 October 1998.

- Warren, W. S., Rabitz, H. & Dahleh, M. Coherent control of quantum dynamics: the dream is alive. *Science* **259**, 1581–1589 (1993).
- Gordon, R. J. & Rice, S. A. Active control of the dynamics of atoms and molecules. *Annu. Rev. Phys. Chem.* **48**, 601–641 (1997).
- Tannor, D. J. & Rice, S. A. Coherent pulse sequence control of product formation in chemical reactions. *Adv. Chem. Phys.* **70**, 441–523 (1998).
- Shapiro, M. & Brumer, P. Coherent and incoherent laser control of photochemical reactions. *Int. Rev. Phys. Chem.* **13**, 187–229 (1994).
- Judson, R. S. & Rabitz, H. Teaching lasers to control molecules. *Phys. Rev. Lett.* **68**, 1500–1503 (1992).
- Bardeen, C. J. et al. Feedback quantum control of molecular electronic population transfer. *Chem. Phys. Lett.* **280**, 151–158 (1997).
- Dupont, E., Corkum, P. B., Liu, H. C., Buchanan, M. & Wasilewski, Z. R. Phase-controlled currents in semiconductors. *Phys. Rev. Lett.* **74**, 3596–2599 (1995).
- Hache, A. et al. Observation of coherently controlled photocurrent in unbiased, bulk GaAs. *Phys. Rev. Lett.* **78**, 306–309 (1997).
- Kleiman, V., Zhu, L., Li, X. & Gordon, R. J. Coherent phase control of the photoionization of H₂S. *J. Chem. Phys.* **102**, 5863–5866 (1995).
- Weinacht, T. C., Ahn, J. & Bucksbaum, P. H. Measurement of the amplitude and phase of a sculpted Rydberg wave packet. *Phys. Rev. Lett.* **80**, 5508–5511 (1998).
- Leichtle, C., Schleich, W. P., Averbukh, I. S. & Shapiro, M. Quantum state holography. *Phys. Rev. Lett.* **80**, 1418–1421 (1998).
- Xin, C. & Yeazell, J. A. Analytical wave-packet design scheme: control of dynamics and creation of exotic wave packets. *Phys. Rev. A* **56**, R2274–R2277 (1997).
- Fittinghoff, D. N. et al. Measurement of the intensity and phase of ultraweak, ultrashort laser pulses. *Opt. Lett.* **21**, 884–886 (1996).
- Gallagher, T. F., Humphrey, L. M., Cooke, W. E., Hill, R. M. & Edelstein, S. A. Field ionization of highly excited states of sodium. *Phys. Rev. A* **16**, 1098–1108 (1977).
- Kinrot, O., Averbukh, I. S. & Prior, Y. Measuring coherence while observing noise. *Phys. Rev. Lett.* **75**, 3822–3825 (1995).
- Hanbury-Brown, R. & Twiss, R. Q. Correlation between photons in two coherent beams of light. *Nature* **175**, 27–29 (1956).
- Schleich, W. & Raymer, M. Quantum state preparation and measurement. *J. Mod. Opt.* **44**, 2021–2022 (1997).
- Freyberger, M., Bardoff, P., Leichtle, C., Schrade, G. & Schleich, W. The art of measuring quantum states. *Phys. World* **10**, 41–45 (1997).
- Leibfried, D., Pfau, T. & Monroe, C. Shadows and mirrors: reconstructing quantum states of atom motion. *Phys. Today* **51**, 22–28 (1998).
- Leonhardt, U. *Measuring the Quantum State of Light* (Cambridge Univ. Press, 1997).
- Strickland, D. & Mourou, G. Compression of amplified chirped optical pulses. *Opt. Commun.* **55**, 447–449 (1985).
- Tull, J. X., Dugan, M. A. & Warren, W. S. High resolution, ultrafast laser pulse shaping and its applications. *Adv. Opt. Magn. Reson.* **20**, 1–51 (1996).
- Raman, C., Conover, C. W. S., Sukenik, C. I. & Bucksbaum, P. H. Ionization of Rydberg wave packets by subpicosecond, half-cycle electromagnetic pulses. *Phys. Rev. Lett.* **76**, 2436–2439 (1996).
- Krause, J. L., Schafer, K. J., Ben-Nun, M. & Wilson, K. R. Creating and detecting shaped Rydberg wave packets. *Phys. Rev. Lett.* **79**, 4978–4981 (1997).

Acknowledgements. We thank D. Tannor, J. Krause and J. Cohen for discussions. This work was supported by the NSF.

Correspondence and requests for materials should be addressed to T.C.W. (e-mail: tweinach@umich.edu).

Nanometre-scale rolling and sliding of carbon nanotubes

M. R. Falvo^{*†}, R. M. Taylor II^{‡‡}, A. Helser[‡], V. Chi[‡],
F. P. Brooks Jr[‡], S. Washburn^{*†} & R. Superfine^{*†}

Departments of ^{*}Physics and Astronomy, and [‡]Computer Science,
[†]North Carolina Center for Nanoscale Materials, University of North Carolina,
Chapel Hill, North Carolina 27599, USA

Understanding the relative motion of objects in contact is essential for controlling macroscopic lubrication and adhesion, for comprehending biological macromolecular interfaces, and for developing submicrometre-scale electromechanical devices^{1,2}. An object undergoing lateral motion while in contact with a second object can either roll or slide. The resulting energy loss and mechanical wear depend largely on which mode of motion occurs. At the macroscopic scale, rolling³ is preferred over sliding, and it is expected to have an equally important role in the microscopic domain. Although progress has been made in our understanding of the dynamics of sliding at the atomic level⁴, we have no comparable insight into rolling owing to a lack of experimental data on microscopic length scales. Here we produce controlled rolling of carbon nanotubes on graphite surfaces using an atomic force microscope. We measure the accompanying energy loss and compare this with sliding. Moreover, by reproducibly rolling a nanotube to expose different faces to the substrate and to an external probe, we are able to study the object over its complete surface.

The microscopic aspects of tribology have been explored by means of the surface force apparatus (SFA)⁵ and atomic force microscopy (AFM)⁶, which have identified the intrinsic dependence of sliding friction on contact area^{7,8} and crystallographic orientation^{9,10}. Further, AFM has been used to perform friction mapping of surfaces with atomic resolution⁴ and has identified stick-slip motion in the sliding of nanometre-scale objects¹¹. Using AFM manipulation^{12,13} of multiwall carbon nanotubes (CNTs), we have observed sliding and rolling. CNTs, with a range of available sizes down to the molecular scale, serve as interesting model systems for tribological studies. In addition, nanotubes are expected to play a part in future nanometre-scale electrical-mechanical devices, and rolling fullerenes have been proposed as ideal lubricants^{14–16}. Our evidence for sliding and rolling comprises sequences of topographical images of manipulated nanotubes and lateral force data acquired during the manipulation to measure energy loss.

Samples were prepared by solvent evaporation on mica and graphite substrates from an ethanol solution of raw carbon soot produced by the carbon arc technique¹⁷. The multiwall CNTs were imaged and manipulated under ambient conditions. The nanomanipulator AFM system^{18,19}, designed for manipulation, comprises an advanced visual interface, teleoperation capabilities for manual control of the AFM tip and haptic (touch) presentation of the AFM data (Topometrix Discoverer). Normal force, lateral force and AFM tip trajectory are recorded simultaneously for strict correlation. The AFM tip is used to apply lateral forces at locations along the tube to produce translations and rotations. These tubes are free of pinning material and move without deformation. The lateral force values have been calibrated from measured cantilever and tip geometry, literature values for the cantilever (Si) elastic moduli and the detector response from the z-translation of the cantilever. We estimate the absolute error in this calibration to be about 30%, with the largest contribution in the uncertainty coming from the AFM tip geometry.

We first summarize our results for manipulation on mica and graphite, then focus on the results relevant to rolling. An extended

object such as a CNT can be pushed from the end or from the side, with possible outcomes including sliding, rolling or some combination of both. On mica, end-on pushes produce a single, pronounced initial stick-slip peak in the lateral force trace which is absent in side-on pushes. Side-on pushes produce no stick-slip features and are accompanied by an in-plane rotation of the nanotube. On graphite, end-on pushes show a series of stick-slip peaks. However, side-on pushes on graphite reveal new behaviour. In addition to cases where a relatively smooth lateral force trace is accompanied by an in-plane rotation (sliding), we also observe cases where the CNT has a pronounced, periodic, stick-slip modulation. In this case, which we refer to as rolling, the tube undergoes no in-plane

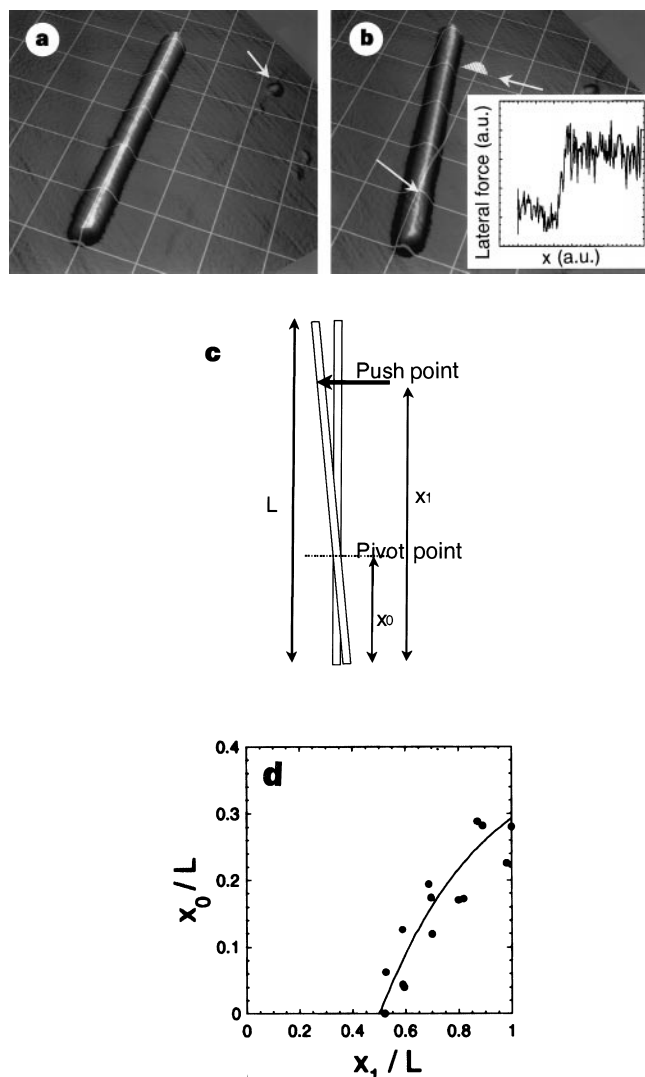


Figure 1 Sliding carbon nanotube. **a**, The tube lies in its original position. Grid lines are overlain so that one of the grid axes corresponds to the original orientation of the tube axis. The absolute position of the grid relative the fiducial feature indicated by the arrow was adjusted to be consistent with that in **b** so that the point of rotation could be determined. **b**, The tube's orientation after AFM manipulation. The pivot point and push point are indicated by the bottom and top arrows, respectively. The white dashes to the right of the push point are markers indicating the trajectory of the AFM tip during manipulation. Inset shows the lateral force trace during a sliding manipulation. The step height is approximately ten times smaller than the stick peaks shown in Fig. 2g. **c**, Diagram of the pushing process. Measuring from the bottom of the tube, x_1 is the push point, and x_0 is the point of rotation (or pivot point). **d**, The relation between the push point and the pivot point for several pushing trials is compared with equation (1) in the text, with no fitting parameters (plotted as a solid curve).

rotation, and the topographical data change in an appropriate fashion for roll-wise reorientation.

We describe the sliding case first. When the nanotube is manipulated from the side on mica or graphite, one class of outcomes features a relatively smooth lateral force jump (inset in Fig. 1b). These manipulations are accompanied by an in-plane rotation of the CNT about a pivot point that depends on the location of the AFM tip during the push. This dependence is shown in Fig. 1, where the AFM images show the position of the AFM tip during the push of a nanotube on graphite, and the overlay of a measuring grid used to determine the tube rotation and the pivot point. A unique relation between the point of manipulation and the point of rotation can be derived with no fitting parameters²⁰ by solving for the motion that minimizes the energy loss due to friction. Uniform friction is assumed, and only sliding/in-plane rotation about an axis perpendicular to the surface is considered. The normalized pivot point location, $x'_0 = x_1/L$, to obtain

$$x'_0 = x'_1 - \sqrt{x'^2_1 - \frac{1}{2}(2x'_1 - 1)} \quad (1)$$

which is compared to the measurements in the inset to Fig. 1b. This approach is easily generalized to other object shapes. This result suggests that the friction between the CNT cylinder and graphite substrate is uniform along the tube and strongly favours the

interpretation of the nanotube motion as sliding. If the nanotube were rolling, the sliding friction at the tube/substrate contact would vary along the length of the tube. We note that this result indicates predictability in the manipulation of nanometer-scale objects, and foreshadows robotic assembly of surface-bound structures^{21,22}.

Most dramatic is the rolling behaviour discovered in the manipulation of large nanotubes from the side-on graphite substrates. In these cases, we have observed changes in nanotube topography correlated with both lateral force stick-slip and the absence of in-plane rotation. The stick-slip signatures in the lateral force data for manipulations in either direction show a periodicity corresponding to the nanotube circumference (Fig. 2g) and are reproducible in detail. The CNT of Fig. 2 (tube A) is 500 nm long and has an average radius at the ends of 13.3 nm and at mid-length of 8.5 nm. The period of the lateral force signals in either direction is 85 ± 2 nm, corresponding to the tube circumference, 83 ± 3 nm. A stick-slip type motion causes the sawtooth shape of these signals. The tip is deflected up to a threshold lateral force and a slip-roll occurs, undeflecting the cantilever. To confirm the stick-slip motion, we performed gradual pushes in which the tip was slowly moved against the CNT while monitoring the lateral force. When the tip was pushed so that the lateral force increased but was turned around before a sudden decrease in the signal (no slip), the subsequently acquired topographical data indicated no resultant movement of the CNT.

The stick-slip peaks were used to reorient reproducibly the CNT on the surface. If the application of the AFM tip was stopped just after a stick-peak, the subsequently acquired topography image showed that the CNT had rotated a fixed amount. The insets between topographs in Fig. 2 show lateral force traces for the pushes between the particular orientations. The change in the tube-end shape from image to image indicates a rolling reorientation. Note that in Fig. 2b, e and 2c, f, tube-end shapes are similar, indicating a similar rolling orientation. In addition to the tube-end shape, the topographical data indicates linear translation between images, which is consistent with a rolling-without-slipping motion. Particularly notable is the upper end of the tube, which has a conical cap consistent with examples identified by transmission electron microscopy²³. A set of data such as these, taken from several sides, can be assembled to provide a three-dimensional reconstruction of a complex object.

The combination of topographical and lateral force data indicates that this tube is undergoing stick-slip rolling motion. There are two important questions. What is the origin of stick-slip in rolling? And what determines whether rolling or sliding will occur in any given translation? The direct correspondence between lateral force features and the nanotube topography implies that adhesion hysteresis is the dominant energy loss mechanism and is responsible for the stick-slip dynamics. A peeling mechanism for the resistance to rolling has been observed for macroscopic rolling experiments²⁴. For example, the amplitudes of the individual stick-slip peaks should be correlated with the varying contact area of the structured tube over the course of its roll. The amplitudes of these peaks can be placed into the context of the van der Waals adhesion of the nanotube at the surface. We can estimate the adhesion force using the van der Waals attraction for the cylinder-on-flat geometry⁵ for a tube (tube B) for which we have taken calibrated lateral force data in rolling. For tube B ($R = 13.5$ nm, $L = 590$ nm) we calculate an adhesive force of 800 nN. This calculation assumes a distance between tube and surface equal to the interlayer spacing in graphite (0.34 nm) and assumes no intervening medium. This force, which completely separates the tube from the substrate, should represent an upper bound for the force required for rolling. In fact, our rolling stick-slip peaks fall well below this value, in the range 20–50 nN.

We can compare our measurements of the sliding friction for a variety of nanotubes by using a simple model for the force due to friction, given by $F = swL$, where s is the interfacial shear stress, w is

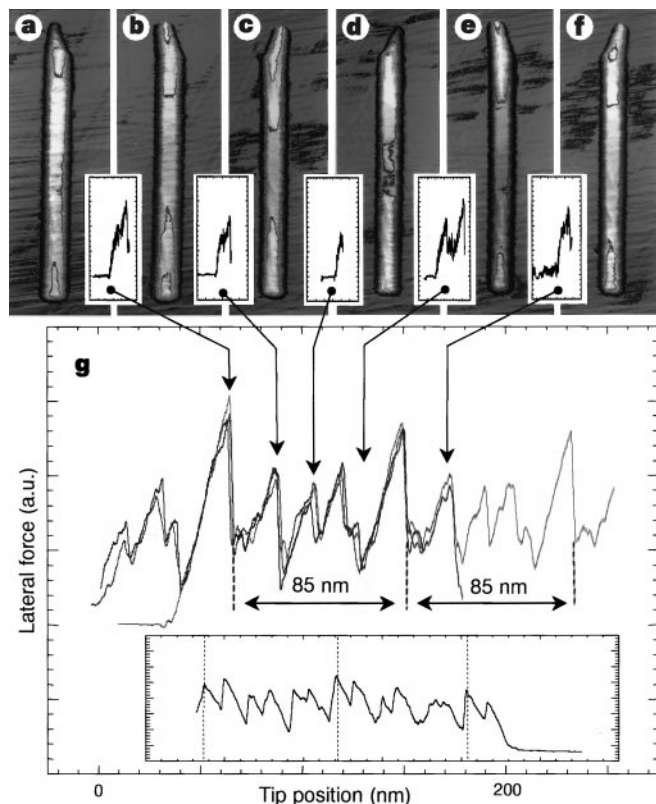


Figure 2 Rolling carbon nanotube. **a–f**, The tube as it is manipulated from left to right. The tube is imaged before and after each of the 5 manipulations. The insets between each topographical image show the lateral force during each manipulation. The tube is moving from left to right not gradually, but in sudden slips in a stick-slip type rolling motion. In **g**, three overlapped signals from separate rolling trials are shown for the lateral force as the tube is pushed through several revolutions of stick-slip rolling motion. The features of the force traces are remarkably reproducible. The 85 nm periodicity in the signal, indicated by the dashed lines, is equal to the circumference of the tube at its ends. A signal for a push in the opposite direction is shown in the inset at the bottom of **g**. This signal differs in form but also has a reproducible 85 nm periodicity. Stick-slip phenomena are known to be hysteretic, explaining the lack of correspondence in the details of the forward and reverse traces.

the contact width, and L is the nanotube length. These quantities are difficult to measure independently and calculations are model-dependent. As a guide, we apply a *JKR* model for the contact of cylinders²⁵ and find a contact width of 3 nm for a tube radius of 13.5 nm. Our measurements of 0.006 N m^{-1} for the friction force per unit length is then consistent with a shear stress of $2 \times 10^6 \text{ Pa}$. This can be compared with a value of $5 \times 10^6 \text{ Pa}$, as inferred from AFM tip/graphite measurements¹⁶. To compare rolling and sliding in a single tube, we can calculate the force (4 nN for $L = 590 \text{ nm}$) and that would be needed to slide tube B, which in fact rolls. Finally, we note that the area under the lateral force trace is a direct measure of energy loss in rolling. For tube B, we measure an energy loss of $(8 \pm 3) \times 10^{-16} \text{ J}$ per revolution. The sliding energy loss expected for this distance (85 nm) can be calculated using the frictional force of 4 nN, yielding $3 \times 10^{-16} \text{ J}$.

When we compare our lateral force measurements for sliding and rolling cases, we find that the stick peaks in rolling are higher than the lateral force needed to sustain sliding, and that the energy cost for rolling is larger than that of the sliding cases. Why should the nanotubes roll? We speculate that, owing to the size and surface features of the rolling nanotubes, a stick peak for sliding in side-on pushing might exist that is larger than the threshold for rolling. Atomic-scale substrate interactions may also play a role as we have observed this characteristic rolling only on graphite. Rolling behaviour has been accompanied by a preferential, threefold, in-plane orientation that indicates intimate nanotube/graphite contact, and perhaps lattice registry. Rolling may occur only when both the nanotube and the underlying graphite have long-range order. In these cases that there may be a barrier for sliding which is larger than that for rolling and may preclude the direct measurement of sliding friction⁷. □

Received 27 July; accepted 10 September 1998.

- Persson, B. N. J. *Sliding Friction: Physical Properties and Applications* (Springer, Berlin, 1998).
- Bhushan, B., Israelachvili, J. N. & Landman, U. Nanotribology: friction, wear and lubrication at the atomic scale. *Nature* **374**, 607–616 (1995).
- Johnson, K. L. *Contact Mechanics* (Cambridge University Press, New York, 1994).
- Mate, M. C., McClelland, G. M., Erlandsson, R. & Chiang, S. Atomic-scale friction of a tungsten tip on a graphite surface. *Phys. Rev. Lett.* **59**, 1942–1945 (1987).
- Israelachvili, J. *Intermolecular and Surface Forces* (Academic, San Diego, 1991).
- Binnig, G., Quate, C. F. & Gerber, C. Atomic force microscope. *Phys. Rev. Lett.* **56**, 930–933 (1986).
- Sheehan, P. E. & Lieber, C. M. Nanotribology and nanofabrication of MoO_3 structures by atomic force microscopy. *Science* **272**, 1158–1161 (1996).
- Johnson, K. L. in *Micro/nanotribology and its Applications* (ed. Bhushan, B.) 157 (Kluwer Academic, Dordrecht, The Netherlands, 1997).
- Hirano, H., Shinjo, K., Kaneko, R. & Murata, Y. Anisotropy of frictional forces in muscovite mica. *Phys. Rev. Lett.* **67**, 2642–2645 (1991).
- Overney, R. M., Takano, H., Fujihira, M., Paulus, W. & Ringsdorf, H. Anisotropy in friction and molecular stick-slip motion. *Phys. Rev. Lett.* **72**, 3546–3549 (1994).
- Luthi, R. *et al.* Sled-type motion on the nanometer scale: Determination of dissipation and cohesive energies of C_{60} . *Science* **266**, 1979–1981 (1994).
- Falvo, M. R. *et al.* Bending and buckling of carbon nanotubes under large strain. *Nature* **389**, 582–584 (1997).
- Hertel, T., Martel, R. & Avouris, P. Manipulation of individual carbon nanotubes and their interactions with surfaces. *J. Phys. Chem. B* **102**, 910–915 (1998).
- Rapoport, L. *et al.* Hollow nanoparticles of WS_2 as potential solid-state lubricants. *Nature* **387**, 791–793 (1997).
- Bhushan, B., Gupta, B. K., Van Cleef, G. W., Capp, C. & Coe, J. V. Sublimed C_{60} films for tribology. *Appl. Phys. Lett.* **62**, 3253–3255 (1993).
- Schwarz, U. D., Zwornier, O., Koster, P. & Wiesendanger, R. Quantitative analysis of the frictional properties of solid materials at low loads. I. Carbon compounds. *Phys. Rev. B* **56**, 6987–6996 (1997).
- Ebbesen, T. W. & Ajayan, P. M. Large-scale synthesis of carbon nanotubes. *Nature* **358**, 220–222 (1992).
- Taylor, R. M. *et al.* SIGGRAPH '93 (ACM SIGGRAPH, New York, 1993).
- Finch, M. *et al.* ACM Symposium on Interactive 3D Graphics (ACM SIGGRAPH, Monterey, California, 1995).
- Prescott, J. *Mechanics of Particles and Rigid Bodies* (Longmans, Green and Co., London, 1936).
- Baur, C. *et al.* Robotic nanomanipulation with a scanning probe microscope in a networked computing environment. *J. Vac. Sci. Tech. B* **15**, 1577–1580 (1997).
- Mason, M. T. & Salisbury, J. K. *J. Robot. Hands and the Mechanics of Manipulation* (MIT Press, Cambridge, Massachusetts, 1985).
- Ebbesen, T. W. & Takada, T. Topological and sp^3 defect structures in nanotubes. *Carbon* **33**, 973–978 (1995).
- Kendall, K. Rolling friction and adhesion between smooth solids. *Wear* **33**, 351–358 (1975).
- Chaudhury, M. J., Weaver, T., Hui, C. Y. & Kramer, E. J. Adhesive contact of cylindrical lens and a flat sheet. *J. Appl. Phys.* **80**, 30–37 (1996).

Acknowledgements. We thank O. Zhou for CNT soot material, S. Paulson for helping to solve sample-preparation problems, and the entire Nanomanipulator team. The NSF, the NIH, the Office of Naval Research, and Topometrix Inc. provided financial support for this work.

Correspondence and requests for materials should be addressed to R.S. (e-mail: rsuper@physics.unc.edu).

Identifying the forces responsible for self-organization of nanostructures at crystal surfaces

K. Pohl*, M. C. Bartelt*, J. de la Figuera*, N. C. Bartelt*, J. Hrbek*† & R. Q. Hwang*

*Sandia National Laboratories, Livermore, California 94550, USA

The spontaneous formation of organized surface structures at nanometre scales^{1,2} has the potential to augment or surpass standard materials patterning technologies. Many observations of self-organization of nanoscale clusters at surfaces have been reported^{1–10}, but the fundamental mechanisms underlying such behaviour—and in particular, the nature of the forces leading to and stabilizing self-organization—are not well understood. The forces between the many-atom units in these structures, with characteristic dimensions of one to tens of nanometres, must extend far beyond the range of typical interatomic interactions. One commonly accepted source of such mesoscale forces is the stress field in the substrate around each unit^{1,11–13}. This, however, has not been confirmed, nor have such interactions been measured directly. Here we identify and measure the ordering forces in a nearly perfect triangular lattice of nanometre-sized vacancy islands that forms when a single monolayer of silver on the ruthenium (0001) surface is exposed to sulphur at room temperature. By using time-resolved scanning tunnelling microscopy to monitor the thermal fluctuations of the centres of mass of the vacancy islands around their final positions in the self-organized lattice, we obtain the elastic constants of the lattice and show that the weak forces responsible for its stability can be quantified. Our results are consistent with general theories of strain-mediated interactions between surface defects in strained films.

We began by depositing slightly less than one monolayer (ML) of Ag at room temperature on a clean Ru(0001) surface prepared in ultrahigh vacuum. A Ag film forms that is highly strained as a result of the ~7% lattice mismatch between Ag and Ru. A flash anneal to 750 K, to obtain an equilibrium Ag film, one single layer high, invariably produces an ordered pattern in the film (unit cell ~40 × 60 Å) consisting of a near-square array of threading dislocations^{14,15}. Subsequent exposure of this strained layer to sulphur, at room temperature, triggers strikingly complex behaviour in the Ag film, as sulphur displaces Ag atoms. Here we describe two main regimes distinguished by the sulphur coverage, θ_s .

For low $\theta_s < 0.05 \text{ ML}$, scanning tunnelling microscopy (STM) images show far-separated two-dimensional vacancy islands of highly regular size, about $34 \pm 11 \text{ Å}$ diameter (Fig. 1a, b). Sulphur is found coating exposed Ru regions, including the inside of the islands where it forms ordered two-dimensional clusters (Fig. 1b inset). This is not surprising because the interaction of sulphur with Ru is considerably stronger than that with Ag, and it certainly reduces the energy cost of exposing the Ru(0001) surface by decreasing its surface energy.

The isolated vacancy islands are extremely mobile (Fig. 1a, b), significantly more so than has been reported for vacancy islands of similar size on clean metal (111) surfaces^{16,17}. Islands diffuse perhaps by atoms diffusing along the edges of islands, although other mechanisms involving a flow of atoms across an island, or exchange of edge atoms with the sulphur and Ag adatom gas, could also be significant. We measured island hop rates of several hundreds of

† Permanent address: Brookhaven National Laboratory, Upton, New York 11973, USA

ångströms per minute, at room temperature, each hop averaging 50 Å between positions seemingly related to the network of dislocations in the original Ag film. This high island mobility suggests that these structures are equilibrated.

As we monitor this dilute 'gas' of islands with the STM, we often observe random formation and dissociation of small collections of islands, with island–island separations of ~ 50 Å (Fig. 1a, b). Strong short-range repulsions between islands seem to prevent them from coming closer to one another and coalescing. The effect of these

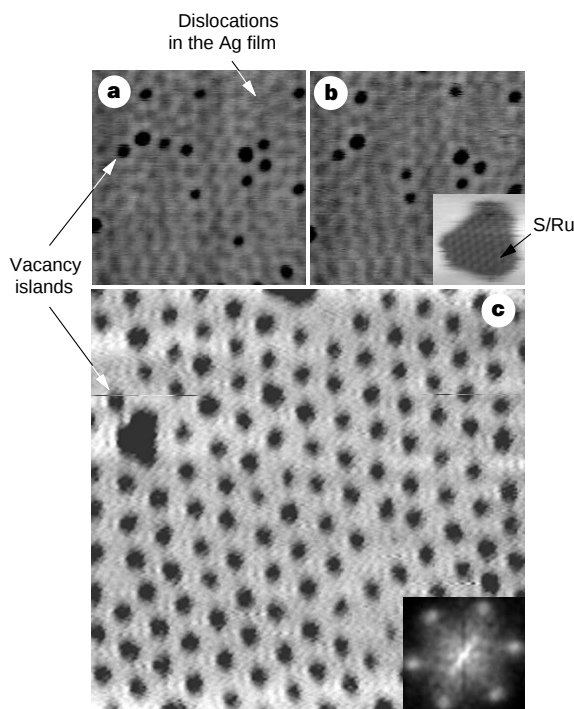


Figure 1 Low- and high-density vacancy island lattices. **a, b**, Two frames, 2 min apart, of an STM 'movie' of the same 480×480 Å area of sulphur-exposed Ag/Ru(0001). The sulphur coverage is $\theta_s \approx 0.04$ ML, and the area fraction of vacancy islands is $f \approx 0.07$. Inset in **b**, 60×60 Å STM image of a large vacancy island (these are immobile compared to smaller islands, but 'breathe' on the timescale of STM image acquisition.) Inside the island is a cluster of sulphur atoms. Diffusing single vacancies are often observed in such clusters, disappearing at the island edge within minutes. We note how the motion of the island edges is so rapid (compared to the rate of STM image acquisition) that they have a 'blurred' appearance. **c**, A 600×600 Å STM image of the vacancy island crystal (here, $\theta_s \approx 0.1$ ML and $f \approx 0.2$). Defects in the crystal include large islands (density $\sim 2 \times 10^{-5}$ Å $^{-2}$) and paired dislocation lines (density $\sim 7 \times 10^{-6}$ Å $^{-2}$). Inset in **c**, the Fourier transform of a 200×200 nm² image.

repulsions becomes increasingly dramatic as we add more sulphur, and the density of vacancy islands, as well as the density and size of self-assembled clusters of islands, increase. Near 0.1 ML of sulphur, a stable, nearly-perfect, triangular lattice of vacancy islands forms, with island–island separations of ~ 53 Å, extending over each of the micrometre-wide Ru terraces (Fig. 1c). The distribution of island diameters in the lattice is also impressively narrow (± 4 Å), peaked at 24 Å. Annealing to temperatures below the desorption temperature of the Ag produces no apparent changes in the lattice.

A large series of consecutive STM images reveals that the islands fluctuate about their final positions in the lattice, probably by the same mechanisms of island diffusion that were apparent in the low-island-density regime. This island motion is illustrated by the image overlay of Fig. 2a. Quantitative analysis of these fluctuations allows direct assessment of the strength of the interactions between islands, as we show next. We note that the individual island areas also fluctuate in time, perhaps via exchange of Ag edge atoms between neighbouring islands or with the adatom gas. However, we did not observe any consistent correlation between fluctuations in the island areas and those in the island positions, not even for nearest-neighbour islands. The relative amplitude of the fluctuations in the island positions is also significantly larger than that of the areas.

For analysis of the lattice fluctuations, we then followed the displacement, $\mathbf{u}$, of the centre of mass (CM) of each island in a 450×450 Å array (~ 80 islands) with time-resolved STM, recording images of the same area at intervals of 10–20 s for nearly 4 hours. We define $\mathbf{u}(\mathbf{r}_0, t) = \mathbf{r}(t) - \mathbf{r}_0$, where $\mathbf{r}(t)$ is the position of the CM at time t of the vacancy island at the equilibrium CM position $\mathbf{r}_0$. The typical range of values for the components of $\mathbf{u}$, obtained by marking the CM trajectories over the complete data set, is shown for one island in Fig. 2b. The motion of individual islands is correlated, reflecting interactions between neighbouring islands. Figure 2c shows clearly that $\mathbf{u}$ -values are positively correlated for neighbouring islands ('in phase' displacements being more likely, as expected for repulsive island–island interactions). This correlation decays rapidly with island–island separation.

As the measured amplitude of the island fluctuations is small, we expect them to be well described as vibrations in a harmonic crystal. (Strictly speaking, here there is no analogue of the inertia in a crystalline solid, as the motion of the islands is purely diffusive.) In this case, each component of the Fourier-transformed CM displacements is expected to fluctuate independently, while satisfying equipartition of energy. This allows a rigorous connection between the amplitude of the fluctuations and the elastic constants of the 2D island crystal^{18,19}. We focus on the long-wavelength behaviour of the expectation value of the steady-state pair correlation function, defined as $G_{ij}(\mathbf{q}) = (1/A)[\langle \hat{u}_i(\mathbf{q})\hat{u}_j(-\mathbf{q}) \rangle]$, between components i and j of the Fourier-transformed displacements, $\hat{\mathbf{u}}(\mathbf{q}) = \int d^2\mathbf{r}_0 \mathbf{u}(\mathbf{r}_0) \exp(i\mathbf{q} \cdot \mathbf{r}_0)$. Here, A is the analysed area. The

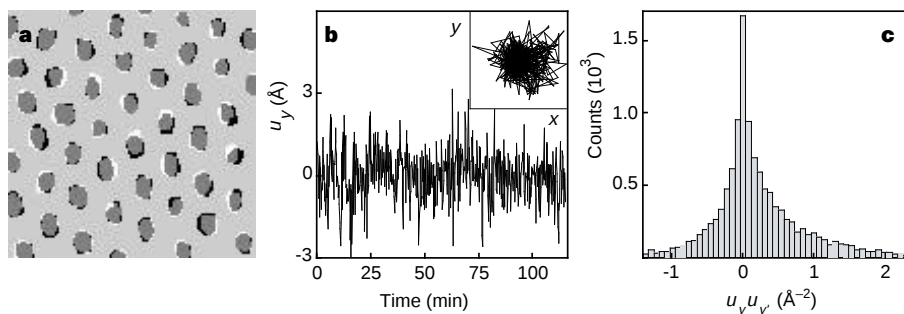


Figure 2 Thermal fluctuations in the vacancy island crystal. **a**, Overlay of two 300×300 Å (contrast enhanced) STM images, 20 s apart. In one of the frames the islands are black and in the other white, so the overlap is dark grey. **b**, The y -component of the CM displacement, u_y , versus time for a selected island, taken from a sequence of 690 frames, 10 s apart. The inset shows the full x - y trajectory.

c, Distribution of the (equal-time) product $u_y u_{y'}$, of CM displacements of nearest-neighbour islands, over the 690 frames. The average is $\langle u_y u_{y'} \rangle \approx 0.26$ Å $^{-2}$. Skewness, $\kappa \approx 1.8$, in this distribution reflects (positive) correlation in the motion of nearest-neighbour islands.

components i and j of $\hat{u}(\mathbf{q})$ can equivalently be distinguished as components which are longitudinal (l) or transverse (t) to $\mathbf{q}$. For small, non-zero $q = |\mathbf{q}|$, equipartition of energy amongst the independent modes then gives^{18,19} $G_{tt}(\mathbf{q}) = k_B T / (q^2 \mu)$ and $G_{ll}(\mathbf{q}) = [k_B T / (q^2 \mu)] \times [1 - (\lambda + \mu) / (\lambda + 2\mu)]$, where T is the temperature, k_B is the Boltzmann constant, and λ and μ are the elastic Lamé coefficients of the 2D lattice which govern changes in the elastic energy density. (The bulk modulus is $\lambda + \mu$; the shear modulus is μ .) Thus, estimates of λ and μ can be obtained by analysing either the slopes of G_{tt} and G_{ll} versus $1/q^2$, or the actual values of G_{tt} and G_{ll} at some (small) q .

For a 'perfectly' ordered 6×6 vacancy island cell, Fig. 3 shows that the amplitude of G_{ll} decreases rapidly with q , qualitatively as predicted. From these and similar data for G_{tt} , we measured average values of $G_{ll} \approx 3 \pm 1 \text{ nm}^4$ and $G_{tt} \approx 4 \pm 2 \text{ nm}^4$, for the smallest non-zero $q = 4\pi / (10\sqrt{3}b) \approx 0.14 \text{ nm}^{-1}$. Thus, from the relations above, we estimate $\lambda \approx \mu \approx 0.02\text{--}0.04 \text{ N m}^{-1}$. The uncertainty reflects observed variation in the values obtained at different q . These results are consistent with values obtained from the width of the gaussian distribution of mean square displacements, or from analysis of the shape and skewness of the distribution of displacement correlations for nearest-neighbour islands in Fig. 2c (K.P. *et al.*, manuscript in preparation).

In order to use the above analysis to elucidate the origin of the forces that stabilize the island vacancy lattice, one must consider two competing contributions. One is from the line tension, γ_b , which reflects the cost of creating island edges, where atoms have reduced coordination. This contribution favours the formation of a few large islands. For Ag(111), $\gamma_b \approx 0.1 \text{ eV } \text{\AA}^{-1}$ (ref. 20). To counter this tendency, there are elastic forces between the edges of neighbouring islands. These forces come from the elastic deformations of atomic positions around the edges of islands, which arise in response to the strain in the film. (Ag has a larger lattice constant than Ru, so a slight outward expansion at the island edges is expected.) The interference between these deformations leads to an interaction that is inversely proportional to the distance between pairs of island edges¹². The proportionality constant is an energy per unit length, γ_d , which characterizes the interaction strength. The sign of this interaction depends on the cosine of the angle between pairs of edges so that neighbouring edges of adjacent islands (that are of opposite orientation) lead to a repulsive force, while edges of the same orientation from adjacent islands are attractive (weaker entropic

and elastic dipole repulsions being neglected)¹³. The net contribution to the elastic energy, relative to that of a uniform film, is negative, favouring spontaneous island formation^{1,11,12}. It is γ_d that we want to measure here. We note that the form of the elastic hamiltonian^{11,12} assumes that the Ru substrate deforms in reaction to the elastic relaxations near the island edges: such deformation of the Ru substrate by strain in the Ag film is consistent with the results in ref. 15.

The complete phase diagram for this model has been analysed¹², providing some context in which to interpret our results. In particular, the model predicts the existence of a critical fraction, $f_c \approx 0.29$, of the surface covered with islands, such that, for $f \leq f_c$, the lowest-energy structure is a triangular lattice of near-circular vacancy islands of narrowly distributed radius, R , and repeat distance, b (but other lattice symmetries can be imposed by the substrate¹⁰). These predictions are in qualitative agreement with our observations in Fig. 1. There is some disagreement on the predicted versus observed values of R and b , and their dependence on f , but this is not surprising for a model which overlooks details of the reaction with sulphur, the subsequent displacement of Ag atoms, or the potential role of the dislocation network in the Ag film. For higher f , other ordered structures are predicted to form^{11,12}, but to test the formation of such structures, and related details of the phase diagram, tailored experimental studies are required (K.P. *et al.*, manuscript in preparation).

To estimate the value of γ_d , we used a connection to the elastic constants, λ and μ . If the energy increase due to an isotropic expansion of the lattice unit cell (at constant f) is stored as elastic energy, then one finds $\gamma_d = 3^{1/2}(\lambda + \mu)b^2 / (\pi R)$ (K.P. *et al.*, manuscript in preparation). For our estimates of λ and μ , this relation gives $\gamma_d \approx 0.15\text{--}0.3 \text{ eV } \text{\AA}^{-1}$, using the measured $b \approx 53 \text{ \AA}$ and $R \approx 12 \text{ \AA}$. To assess the feasibility of the elastic forces as the root of stable self-organized behaviour, it is also useful to relate γ_d directly to the stress field around the islands. Continuum elasticity considerations¹³ obtain $\gamma_d \approx (1 - \nu)\sigma^2 / (2\pi\epsilon)$, where ν and ϵ are the Poisson ratio and bulk modulus of the substrate, respectively, and σ is the difference in the surface stress between Ag and Ru regions. Assuming $\nu_{\text{Ru}} \approx 0.5$ and $\epsilon_{\text{Ru}} \approx 10^{10} \text{ N m}^{-2}$, the estimate of γ_d above implies $\sigma \approx 0.3\text{--}0.5 \text{ eV } \text{\AA}^{-1}$. This is within the range of σ values reported for unreconstructed metal(111) surfaces¹³, providing confidence in our central estimate of γ_d .

The S/Ag/Ru(0001) lattice of vacancy islands is unlikely to be unique: small changes in the type of adsorbate or host film will probably produce new lattices with different parameters (R , b , λ , μ and so on). A possible application of vacancy island lattices is their use as templates for pre-patterning new nanostructures¹⁰, thus allowing properties which are unique to these length scales to be controlled. This relies on the detailed characterization, as reported here, of the forces responsible for the existence of such templates. □

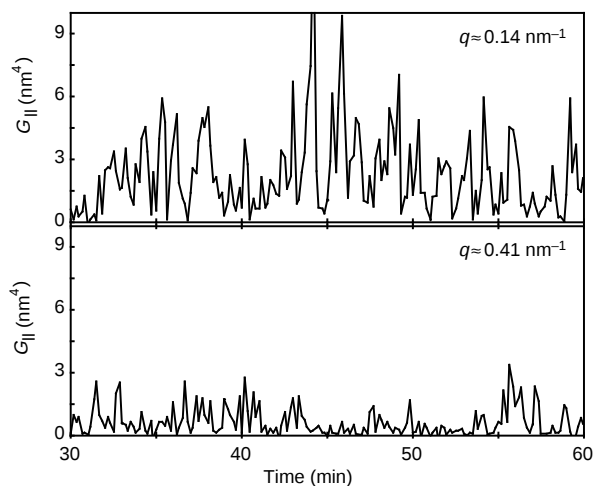


Figure 3 Analysis of correlations in the thermal displacements of neighbouring islands. The time dependence of the longitudinal pair correlation function $G_{ll}(q)$ is shown for $q \approx 0.14 \text{ nm}^{-1}$ and 0.41 nm^{-1} , over 200 STM frames ($\sim 1 \text{ h}$). No apparent correlations in $G_{ll}(q)$, or $G_{tt}(q)$, for different q values were observed. The average values of G_{ll} are 3 nm^4 and 0.7 nm^4 (with larger uncertainty in the latter), for $q \approx 0.14 \text{ nm}^{-1}$ and 0.41 nm^{-1} , respectively.

Received 1 June; accepted 5 October 1998.

- Seul, M. & Andelman, D. Domain shapes and patterns: the phenomenology of modulated phases. *Science* **267**, 476–483 (1995).
- Itano, W. M. *et al.* Bragg diffraction from crystallized ion plasmas. *Science* **279**, 686–689 (1998).
- Evans, J. H. Observations of a regular void array in high purity molybdenum irradiated with 2 MeV nitrogen ions. *Nature* **229**, 403–404 (1971).
- Asai, M., Ueba, H. & Tatsuyama, C. Heteroepitaxial growth of Ge films on the Si(100)- 2×1 surface. *J. Appl. Phys.* **58**, 2577–2583 (1985).
- Kern, K. *et al.* Long-range spatial self-organization in the adsorbate-induced restructuring of surfaces: Cu(110)-(2 $\times$ 1)O. *Phys. Rev. Lett.* **67**, 855–858 (1991).
- Hörnig, H., West, J. R., Conrad, E. H. & Ellialtıoglu, R. Island ordering on clean Pd(110). *Phys. Rev. B* **47**, 13055–13058 (1993).
- Zeppenfeld, P., Krzyzowski, M. A., Romanczyk, Ch., David, R. & Comsa, G. Stability of disk and stripe patterns of nanostructures at surfaces. *Surf. Sci.* **342**, L1131–L1136 (1995).
- Haasen, P. *Physical Metallurgy* 256 (Cambridge Univ. Press, New York, 1996).
- Tersoff, J., Teichert, C. & Lagally, M. G. Self-organization in growth of quantum dot superlattices. *Phys. Rev. Lett.* **76**, 1675–1678 (1996).
- Parker, T. M., Wilson, L. K., Condon, N. G. & Leible, F. M. Epitaxy controlled by self-assembled nanometer-scale structures. *Phys. Rev. B* **56**, 6458–6461 (1997).
- Vanderbilt, D. Phase segregation and work-function variations on metal surfaces: spontaneous formation of periodic domain structures. *Surf. Sci.* **268**, L300–L304 (1992).
- Ng, K.-O. & Vanderbilt, D. Stability of periodic domain structures in a two-dimensional dipolar model. *Phys. Rev. B* **52**, 2177–2183 (1995).
- Ibach, H. The role of surface stress in reconstruction, epitaxial growth and stabilization of mesoscopic structures. *Surf. Sci. Rep.* **29**, 193–263 (1997).

14. Stevens, J. L. & Hwang, R. Q. Strain stabilized alloying of immiscible metals in thin films. *Phys. Rev. Lett.* **74**, 2078–2081 (1995).
15. Hwang, R. Q., Hamilton, J. C., Stevens, J. L. & Foiles, S. M. Near-surface buckling in strained metal overlayer systems. *Phys. Rev. Lett.* **75**, 4242–4245 (1995).
16. Morgenstern, K., Rosenfeld, G., Poelsema, B. & Comsa, G. Brownian motion of vacancy islands on Ag(111). *Phys. Rev. Lett.* **74**, 2058–2061 (1995).
17. de la Figuera, J., Prieto, J. E., Ocal, C. & Miranda, R. Creation and motion of vacancy islands on solid surfaces: a direct view. *Solid State Commun.* **89**, 815–818 (1994).
18. Nelson, D. R. & Halperin, B. I. Dislocation-mediated melting in two dimensions. *Phys. Rev. B* **19**, 2457–2484 (1979).
19. Huse, D. A. Shape of diffraction peaks of a two-dimensional solid. *Phys. Rev. B* **28**, 6110–6111 (1983).
20. Morgenstern, K., Rosenfeld, G. & Comsa, G. Decay of two-dimensional Ag islands on Ag(111). *Phys. Rev. Lett.* **76**, 2113–2116 (1996).

Acknowledgements. This work was supported by the Office of Basic Energy Sciences of the US DOE, Divisions of Materials Science and Chemical Science. J.d.I.F. also acknowledges support from the Fulbright Commission and the Spanish MEC.

Correspondence and requests for materials should be addressed to R.Q.H. (e-mail: rqhwan@california.sandia.gov).

Coordinated proton tunnelling in a cyclic network of four hydrogen bonds in the solid state

Dermot F. Brougham*, Roberto Caciuffo† & Anthony J. Horsewill*

* School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK

† Istituto Nazionale per la Fisica della Materia, Università di Ancona, Dipartimento di Scienze dei Materiali e della Terra, Via Brece Bianche, I-60131 Ancona, Italy

The transfer of protons involved in hydrogen bonding is fundamental to many chemical and biological processes. Quantum tunnelling can play an important role in this process^{1,2}. It manifests itself in strong isotope effects^{3,4} and has been observed directly in the solid state⁵. The tunnelling behaviour seen in such studies usually displays the characteristics of a particle confined in a double-well potential. But proton tunnelling can also occur in a coordinated fashion that involves many hydrogen bonds simultaneously. Such a process may significantly affect the

properties of linear and circular networks of hydrogen bonds, which occur in ice and in macromolecules containing hydroxyl groups^{6,7}. Here we report the direct observation by NMR relaxometry of coordinated proton tunnelling in a cyclic array of four hydrogen bonds in solid *p*-tert-butyl calix[4]arene at low temperature. We are able to quantify the parameters that describe this phenomenon and find good agreement with theoretical predictions for phonon-assisted tunnelling⁸.

Calix[*n*]arenes are bowl-shaped synthetic macrocycles with ordered arrays of hydrogen-bonded phenol-methylene oligomers⁹. The number of the phenolic units can vary from *n* = 3 to *n* = 14, and different substituents can be attached to the aromatic ring in the *para* position and at the phenolic oxygen. The *p*-tert-butyl calix[4]arene molecule possesses four-fold symmetry and assumes a shape (shown in Fig. 1a) that is maintained by a cyclic array of four hydrogen bonds¹⁰.

NMR relaxometry is a well-established experimental technique for the study of molecular dynamics. The molecular motion modulates the nuclear magnetic interactions in which the nucleus participates, and hence drives the transitions that mediate the establishment of equilibrium amongst the nuclear spin energy levels. Therefore, information on the rates and frequencies that characterize the molecular motion is encoded within the nuclear spin relaxation properties.

In the system studied here, the modulation of the nuclear dipole-dipole interactions dominates the proton spin relaxation. In the case where the dynamics are stochastic, the time correlation function for the proton position is an exponential decay, and its Fourier transform, which determines the spin-lattice relaxation rate $T_1^{-1}(\omega)$, has lorentzian form in the frequency domain. As single-spin and double-spin flip transitions are possible, T_1^{-1} samples the motional spectrum at one and two times the proton Larmor frequency, ω_L . Measurements of T_1^{-1} as a function of an external magnetic field, *B*, will therefore give a direct measurement of the spectral density function because $\omega = \gamma B$ (where the magnetogyric ratio $\gamma = 2.675 \times 10^8 \text{ s}^{-1} \text{ T}^{-1}$). In the case where the molecular system interchanges between two tautomeric configurations, labelled *a* and *b*, then for a powdered material¹¹

$$T_1^{-1}(\omega_L) = Cp_a p_b \left[\frac{\tau_c}{1 + \omega_L^2 \tau_c^2} + \frac{4\tau_c}{1 + (2\omega_L)^2 \tau_c^2} \right] \quad (1)$$

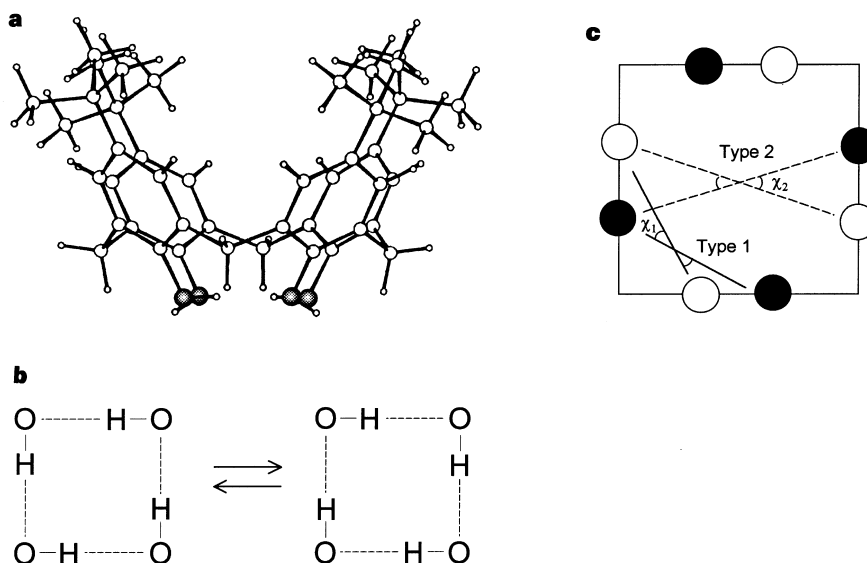


Figure 1 Coordinated proton tunnelling in *p*-tert-butyl calix[4]arene. A side view of the molecular conformation of this compound, perpendicular to the four-fold symmetry axis, is shown in **a**. The oxygen atoms participating in the cyclic hydrogen bond array are shaded. The interconversion between the two tautomers

schematically shown in **b** can be achieved by coordinated proton tunnelling. The relevant geometrical parameters for determining the motion-induced changes in the proton dipolar interactions are defined in **c**, where the hydrogen atoms are indicated by open circles for tautomer *a*, and filled circles for tautomer *b*.

where p_a and p_b are the equilibrium populations of the two tautomers, τ_c is the correlation time, and C is the dipolar constant. C is determined by the changes in molecular geometry that occur with the motion. Therefore, the frequency dependence of the inverse nuclear spin–lattice relaxation time is a superposition of two lorentzian lineshapes with half-widths in the ratio 1:2 and amplitudes in the ratio 4:1. The half-width at half-maximum amplitude is a measure of the tautomer conversion rate, τ_c^{-1} .

Using the saturation-recovery technique in pulsed NMR, incorporating magnetic field-cycling, the frequency dependence of the proton spin–lattice relaxation time in *p*-tert-butyl calix[4]arene was determined following the procedure described in ref. 12. In the temperature range studied, the spin–lattice relaxation exhibited bi-exponential behaviour. This may reflect incomplete averaging by spin diffusion, and is not unusual in complex systems in the solid state. The fast and slow fractions that characterized the magnetization recovery displayed parallel behaviour, and the average spin–lattice relaxation time was represented by a single parameter corresponding to the time when the magnetization recovered to a fraction $(1 - 1/e)$ of its equilibrium value.

The switching time of the magnet used in the field-cycling NMR experiments imposed a limit on the maximum value of T_1^{-1} that could be measured, and hence on the temperature range over which the dynamics could be studied. Measurements of $T_1^{-1}(\omega)$ were made at 15.0, 17.0, 19.0 and 21.0 K and the results are presented in Fig. 2. Only the modulus of frequency is accessible experimentally so, as a visual aid, the data have been reflected in the zero-frequency axis. In this low-temperature region, and over the frequency range measured, the motions of the tert-butyl groups are known to be insignificant so the relaxation is dominated by the proton motion in the cyclic array of hydrogen bonds. At each temperature, the data have been fitted to the superposition of two lorentzians defined by equation (1); the best fits are illustrated by the solid lines in Fig. 2. The data are well represented by the lorentzian functions centred at zero frequency, so we conclude that the correlation functions that define the molecular motion are indeed exponential decays.

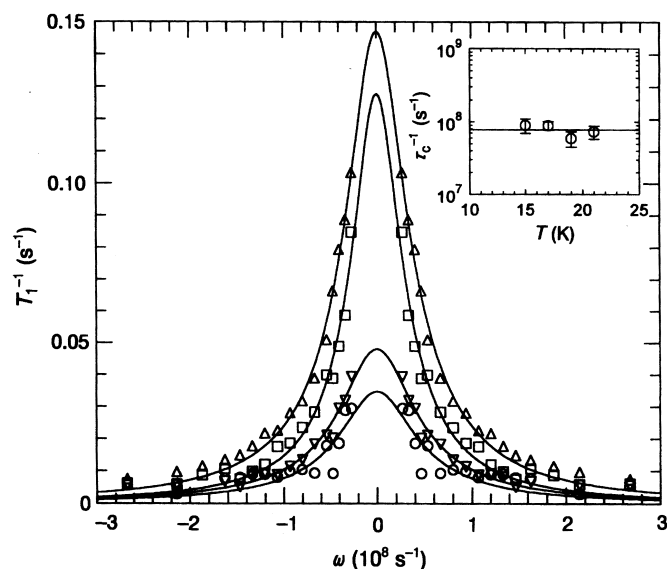


Figure 2 The inverse of the proton spin–lattice relaxation time (T_1^{-1}) as a function of the angular frequency ω , at selected temperatures T . (Circles, 15 K; downward-pointing triangles, 17 K; squares 19 K; upward-pointing triangles, 21 K.) The solid lines are the fit to the spectral density corresponding to an exponential time correlation function, equation (1). The half-width, giving a direct measure of the proton transfer rate, is plotted in the inset. The temperature independence of this quantity is a signature of quantum-mechanical motion.

The inset in Fig. 2 shows the tautomer conversion rates, obtained from the widths of the peaks at half-maximum amplitude, as a function of temperature. As may be seen, this parameter is independent of temperature over the range of experimental conditions ($\tau_c^{-1} = (8 \pm 1) \times 10^7 \text{ s}^{-1}$), thus clearly demonstrating that the dynamical behaviour of the system is non-Arrhenius in character. Classical behaviour would result in a narrowing of the lorentzian lineshapes with decreasing temperature. The observed temperature-independence of τ_c^{-1} is thus the signature of tunnelling, which we attribute to the inter-tautomer conversion schematically shown in Fig. 1b.

The amplitude of the lorentzian lineshapes, shown in Fig. 3 as a function of $1/T$, contains information on the energy difference, A , between the two tautomers. If we assume Boltzmann populations, then $p_a p_b = \text{sech}^2(A/2k_B T)$ (ref. 11). The solid line in Fig. 3 is the fit to this function with $A/k_B = 75 \pm 6 \text{ K}$. Confirmation of this value, over a wider range of temperature T , is provided by the slope of the $\ln(T_1)$ versus $1/T$ curve⁸, obtained from measurements conducted at constant frequency, $\omega = 26 \text{ MHz}$. (See Fig. 3 inset; the dashed line is the fit with slope $A/k_B = 76 \pm 6 \text{ K}$.)

The measured values of the asymmetry energy and the temperature-independent τ_c^{-1} are similar to the dynamics of proton transfer in the hydrogen bonds of benzoic acid¹², but in the present case the interchange between the two tautomers is mediated by a coordinated four-proton transfer in the four hydrogen bonds. As the two tautomeric configurations have different energy, the tunnelling system must interact strongly with the phonon bath. Interaction with phonons gives rise to dissipation, and also leads to phonon-assisted tunnelling. This is a stochastic process, and it is the wave-like nature of the protons that permits penetration of the potential barrier. In the low-temperature limit, the measured inverse correlation time determines the tunnelling rate, $k_0 = (8 \pm 1) \times 10^7 \text{ s}^{-1}$.

A quantitative description of the experimental results can be obtained in the framework of a simple model, as follows. Considering the modulation in the dipolar interactions that arises from the

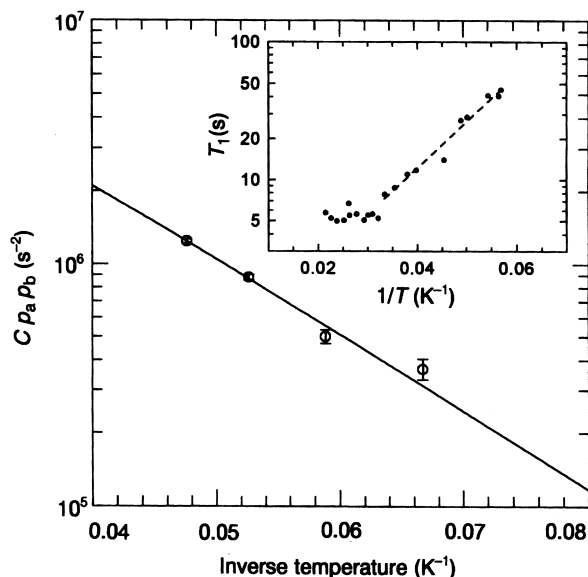


Figure 3 The temperature-dependence of the spectral density amplitude, $C p_a p_b$, for the proton transfer motion. The energy asymmetry between the two tautomers is determined from the slope of this curve ($A/k_B = 75 \pm 6 \text{ K}$). The intercept is a measure of the change in the dipolar interactions arising from the proton motion; the experimental value compares favourably with theoretical estimate, given the known geometry. Confirmation is provided by the temperature-dependence measurements of T_1 at constant frequency, shown in the inset (see text).

hydrogen motion, adapting Andrew and Latanowicz¹¹, we derive the following theoretical expression for the dipolar constant C :

$$C = \frac{1}{N} \sum_{i>j} \frac{9}{10} \gamma^4 \hbar^2 \frac{\sin^2 \chi_{ij}}{r_{ij}^6} \left(\frac{\mu_0}{4\pi} \right)^2 \quad (2)$$

where the vacuum permeability $\mu_0 = 4\pi \times 10^{-7} \text{ H m}^{-1}$, and the indices i, j label the proton pairs. In deriving equation (2), we make the reasonable assumption that a uniform spin temperature is established for all protons in the sample via rapid spin diffusion; $N = 56$ is the total number of protons in the *p*-tert-butyl calix[4]-arene molecule. The geometric parameters χ_{ij} and r_{ij} are determined by the inter-proton vectors in the two tautomers; the angle χ is illustrated in Fig. 1c, and r is the inter-proton distance. We shall assume that all four hydrogen bonds are equivalent, and that the structure approximates to a square as suggested by X-ray diffraction¹⁰. There are six terms in the summation (equation (2)): four equivalent terms relating to internuclear interactions of type 1 (adjacent hydrogen bonds; Fig. 1c), and two equivalent terms of type 2 (hydrogen bonds on opposite sides of the square). Given $r_{\text{O-H...O}} = 2.69 \text{ \AA}$, and making the reasonable assumption that $r_{\text{O-H}} = 1.0 \text{ \AA}$, then for type 1, $\chi_1 = 28.8^\circ$ and $r_1 = 1.96 \text{ \AA}$, and for type 2, $\chi_2 = 28.8^\circ$ and $r_2 = 2.78 \text{ \AA}$. Inserting these values in equation (2) we obtain $C_{\text{calc}} = 1.55 \times 10^8 \text{ s}^{-2}$. The agreement with the observed value, $C_{\text{expt}} = (8 \pm 3) \times 10^7 \text{ s}^{-2}$ that is determined from the fit to the amplitude data in Fig. 3, is as good as one can expect for this model.

Further confirmation is obtained from the depth of the minimum, $(T_1)_{\text{min}}$, revealed in the temperature dependence of T_1 at fixed frequency (Fig. 3 inset). At the higher temperature that characterizes the minimum, the proton transfer rate has increased and $\omega\tau_c = 0.616$. Inserting this condition in equation (1) we obtain, for the spectrometer frequency of 26 MHz:

$$C = (1.15 \times 10^8) / \{(T_1)_{\text{min}} P_a P_b\} \quad (3)$$

Substituting the known values gives $C = 1.16 \times 10^8 \text{ s}^{-2}$, in good correspondence with C_{calc} and C_{expt} above.

An estimate of the height, V_b , of the potential barrier that characterizes the proton transfer dynamics may be obtained by applying the theory of phonon-assisted tunnelling⁸. The low-temperature tunnelling rate, k_0 , is proportional to the square of the tunnelling matrix element, Δ , that characterizes the double minimum potential. With an estimate for the proportionality constant based on the properties of the analogous benzoic acid system⁸, and substituting our experimental value for the tunnelling rate, $k_0 = 8 \times 10^7 \text{ s}^{-1}$, we obtain $\Delta \approx 5 \text{ GHz}$. The tunnelling matrix element depends exponentially on the properties of the barrier; applying a simple one-dimensional description of the double minimum potential¹³, and substituting the value Δ , we obtain $V_b \approx 25 \text{ kJ mol}^{-1}$. This value is consistent with a coordinated four proton transfer process. It is inconsistent with any conceivable sequential or stepwise rearrangement of the protons, a mechanism that in this case would have a barrier height at least one order of magnitude higher. A tunnelling trajectory with such a high barrier would have an exceedingly small probability compared with the coordinated four-proton transfer trajectory across the potential-energy surface.

The agreement between the experimental and calculated values of C , together with the close similarities to the dynamics of benzoic acid, leads us to conclude that we are indeed observing the coordinated proton transfer in the ring of four hydrogen bonds. The dynamics are consistent with the theory of phonon-assisted tunnelling⁸. □

Received 7 April; accepted 20 October 1998.

1. Liu, K. *et al.* Characterisation of a cage form of the water hexamer. *Nature* **381**, 501–503 (1996).
2. MacGillivray, L. R. & Atwood, J. L. A chiral spherical molecular assembly held together by 60 hydrogen bonds. *Nature* **389**, 469–472 (1997).

3. Bahnson, B. J. *et al.* A link between protein structure and enzyme catalyzed hydrogen tunneling. *Proc. Natl Acad. Sci. USA* **94**, 12797–12802 (1997).
4. Aguilar-Parrilla, F., Klein, O., Elguero, J. & Limbach, H. H. A dynamic ¹⁵N NMR study of kinetic hydrogen/deuterium isotope and tunnel effects on the triple proton transfer in crystalline 3,5-dimethylpyrazole. *Ber. Bunsenges. Phys. Chem.* **101**, 889–901 (1997).
5. Oppenländer, A., Rambaud, C., Trommsdorff, H. P. & Vial, J. C. Translational tunnelling of protons in benzoic-acid crystals. *Phys. Rev. Lett.* **63**, 1432–1435 (1989).
6. Saenger, W. Circular hydrogen bonds. *Nature* **279**, 343–344 (1979).
7. Saenger, W., Betzel, Ch., Hingerty, B. & Brown, G. M. Flip-flop hydrogen bonding in a partially disordered system. *Nature* **296**, 581–583 (1982).
8. Skinner, J. L. & Trommsdorff, H. P. Proton transfer in benzoic acid crystals: A chemical spin-boson problem. Theoretical analysis of nuclear magnetic resonance, neutron scattering, and optical experiments. *J. Chem. Phys.* **89**, 897–907 (1988).
9. Diamond, D. & McKervey, M. A. Calixarene-based sensing agents. *Chem. Soc. Rev.* **25**, 15–24 (1996).
10. Andreotti, G. D., Ungaro, R. & Pochini, A. Crystal and molecular structure of cyclo[quarter](5-*t*-butyl-2-hydroxy-1,3-phenylene)methylene]toluene (1:1) clathrate. *J. Chem. Soc. Chem. Commun.* 1005–1007 (1979).
11. Andrew, E. R. & Latanowicz, L. Solid-state proton transfer dynamics and the proton NMR second moment and proton relaxation rates. *J. Magn. Reson.* **68**, 232–239 (1986).
12. Brougham, D. F., Horsewill, A. J. & Jenkinson, R. I. Proton transfer dynamics in the hydrogen bond: a direct measurement of the incoherent tunnelling rate by NMR and the quantum-to-classical transition. *Chem. Phys. Lett.* **272**, 69–74 (1997).
13. Landau, L. D. & Lifshitz, E. M. *Quantum Mechanics* 177 (Pergamon, Oxford, 1958).

Acknowledgements. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) and the Istituto Nazionale per la Fisica della Materia.

Correspondence and requests for materials should be addressed to A.J.H. (e-mail: a.horsewill@nottingham.ac.uk).

Possible predictability in overflow from the Denmark Strait

Bob Dickson*, Jens Meincke†, Ian Vassie‡, Johann Jungclauss & Svein Østerhus||

* Centre for Environment, Fisheries and Aquaculture Science, Lowestoft, Suffolk NR33 0HT, UK

† Institut für Meereskunde, Universität Hamburg, D-22529 Hamburg, Germany

‡ Proudman Oceanographic Laboratory, Bidston Observatory, Merseyside L43 7RA, UK

§ Institut für Meereskunde, Universität Kiel, 24105 Kiel, Germany

|| Geophysical Institute, University of Bergen, N-5007 Bergen, Norway

The overflow and descent of cold dense water from the Denmark Strait sill—a submarine passage between Greenland and Iceland—is a principal means by which the deep ocean is ventilated, and is an important element in the global thermohaline circulation. Previous investigations of its variability—in particular, direct current measurements^{1,2} in the overflow core since 1986—have shown surprisingly little evidence of long-term changes in flow speed. Here we report significant changes in the overflow characteristics during the winter of 1996–97, measured using two current-meter moorings and an inverted echo sounder located at different depths in the fastest part of the flow. The overflow warmed to the highest monthly value yet recorded (2.4 °C), and showed a pronounced slowing and thinning at its lower margin. We believe that the extreme warmth of the overflow caused it to run higher on the continental slope off east Greenland, so that the lower current meters and the echo sounder were temporarily outside and deeper than the fast-flowing core; model simulations appear to confirm this interpretation. We suggest that the extreme warmth of the overflow is a lagged response to a warming upstream in the Fram Strait three years earlier (caused by an exceptional amplification of the winter North Atlantic Oscillation). If this is so, overflow characteristics may be predictable.

Long but non-continuous direct current measurements in the core of the Denmark Strait overflow off southeast Greenland since 1986^{1,2} describe a vigorous, topographically steered, near-bottom current, some 300 m thick, with maximum speeds up to 100 cm s⁻¹, and a mean speed of ~25 cm s⁻¹; there is a large-amplitude fluctuating component with dominant timescales of 2–12 d, but

surprisingly little evidence of variability over longer periods (months, seasons, years).

With the location of the overflow core off Angmagssalik, south-east Greenland, apparently well defined by more than 60 long-term current-meter records², more recent attempts to extend this series to examine decadal timescales have relied on a pair of moorings, three instruments per mooring, set 32 km apart in the fastest part of the flow, in water depths of 2000 m and 2300 m, respectively. As overflow transport is a function of both its speed and its thickness, a bottom-mounted inverted echo sounder (operating at 12 kHz) has been deployed between the two moorings at 2,125 m depth since August 1996. Inverted echo sounders measure the two-way travel time of an acoustic pulse travelling between the sea bed and the surface; but in this application, the sounder was modified to collect and store the acoustic return signals from each point in the overlying water column in order to measure changes in the thickness of the near-bottom current layer. All moorings were recovered by FS *Meteor* in August 1997, providing (to our knowledge) the first simultaneous measurements of overflow speed and thickness.

Both sets of records proved surprising. Current speeds slowed to a few centimetres per second in part of the overflow core which has hitherto been characterized by a vigorous and (when time-averaged over weeks to months) rather steady mean flow. This dramatic change is best shown by plotting the whole current-meter data set since 1986 as monthly means at each mooring site (Fig. 1). All values are from the near-bottom layer (heights of 20, 60 and 100 m from the sea bed), and while exercise-mean speeds ($23\text{--}25\text{ cm s}^{-1}$) at the upper site were close to normal in both of the last two deployments, the mean speeds at the lower mooring began to decrease below normal in January 1997 and dwindled to a minimum of $1\text{--}7\text{ cm s}^{-1}$

in February, according to instrument. At the same time, the monthly average *in situ* temperatures recorded on both moorings rose to an all-time maximum (2.40°C at a height of 100 m from the sea bed, at 2,323 m depth).

Corresponding changes were recorded by the inverted echo sounder. After removing the barotropic (depth-independent) part of the signal, the component of acoustic two-way travel time that was coherent with the bottom temperature record was extracted and converted to give two-hourly estimates of layer thickness for the full year (Fig. 2; salinity and pressure are assumed constant). Though the final temperature corrections have still to be made, the main result is unambiguous—that at the time when the Denmark Strait Overflow Water (DSOW) warmed to its highest recorded temperature in January 1997 and slowed to a minimum at the lower mooring, the near-bottom overflow layer thinned by 80% of its mean thickness (from $\sim 300\text{ m}$ to a minimum of 50 m).

The simplest way of explaining these changes is to assume that as the overflow warmed, the axis of the descending current moved to lie higher on the east Greenland continental slope. In consequence, the upper mooring would remain in the flow and continue to record vigorous mean speeds as normal, but the lower mooring would now lie outside the core and the layer thickness would appear to decrease as the tapering edge of the overflow passed over the echo sounder. We note that in this interpretation, we invoke no long-period changes in the speed of the overflow at its core (Fig. 1a), but merely a translation of the axis of overflow up- and down-slope.

This idea was tested using a hydrostatic, reduced gravity, two-dimensional, primitive equation model which was developed specifically for the simulation of dense overflow plumes such as that in the Denmark Strait, and is capable of resolving the plume horizontally³. A simple experiment in which the sill temperature was raised by 1°C (from -0.3°C to 0.7°C) while salinities were held unchanged (at 34.90) produced a pronounced westward (up-slope) shift at a range downstream equivalent to that of the Angmagssalik array (Fig. 3a). The temperature minimum moves westwards by 15–20 km, but both the layer thickness and temperature show a higher sensitivity to the imposed changes in the down-slope part of the overflow plume (Figs 3b, c). There, at a water depth of 2,400 m, a warming of 0.5°C is accompanied by a 50% decrease in layer thickness; and although along-stream velocities in the core decrease by $<5\text{ cm s}^{-1}$, they slow near this lower edge by up to 10 cm s^{-1} (not shown)—all much as observed.

If these changes all stem basically from the relative warmth of the overflow in 1997, the questions arise as to its cause and why similar changes have not been seen before. The recent extreme behaviour of the North Atlantic Oscillation (NAO) may provide the answer. The NAO is the dominant recurrent mode of atmospheric behaviour in

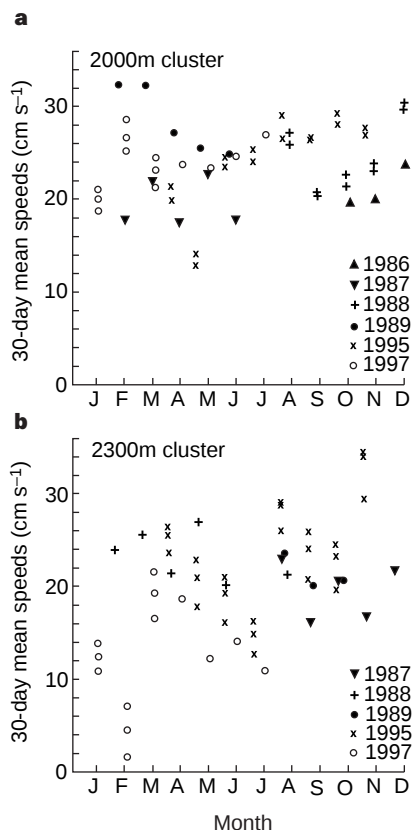


Figure 1 Variability in speed at the core of the Denmark Strait overflow since 1986. Shown are mean current speeds by month for all MAFF current-meter deployments in the near-bottom layers (20–100 m above sea bed) of the Denmark Strait overflow off Angmagssalik, southeast Greenland, at depths of 2,000 m (a) and 2,300 m (b).

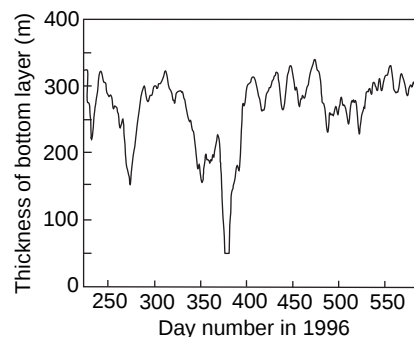


Figure 2 Recent variations in the thickness of the Denmark Strait overflow. Shown are two-hourly estimates of the thickness of the Denmark Strait overflow between August 1996 and August 1997, measured acoustically using a bottom-mounted 12-kHz inverted echo sounder located in the overflow plume off southeast Greenland at $63^\circ 22' \text{ N}$, $36^\circ 04' \text{ W}$ in 2,125 m water depth.

the North Atlantic sector^{4,5} and is one of the most robust on Earth⁶. Over the period of the instrumental record (since 1865) it has exhibited considerable long-term variability⁷, amplifying to its most persistent and extreme positive phase in the late 1980s/early 1990s. As a result, the temperature of the Atlantic water flowing north to the Arctic through Fram Strait increased to its highest recorded values in the early 1990s (Fig. 4a, b; ref. 8). Plainly, to establish a causal link between the extreme warmth in both areas we must first establish a physical connection. First, there is hydrographic evidence that Norwegian Atlantic Water (NAW) makes a sizeable contribution to DSOW. We note that this possibility is only considered in the more recent literature. The first detailed studies of the origins of DSOW^{9,10} restricted its source to the Arctic domain of the Nordic seas as the product of intermediate-depth convection. Only later^{11,12} were DSOW characteristics linked to the current which flows south along the western boundary of the Nordic seas, carrying Atlantic, Polar and Arctic waters towards the Denmark Strait at intermediate depths. Various suggestions have since been made regarding the contribution which this boundary current might make to DSOW formation. According to Mauritzen¹³, DSOW characteristics are formed from 85% of cooled NAW mixed with only 15% of waters

from the convective gyres in the Greenland and Iceland Seas; the NAW is contributed both from the Atlantic waters which recirculate westwards within Fram Strait and those which make a more extended loop through part of the Arctic Ocean. Others¹⁴ suggest that DSOW is formed from equal proportions of NAW from Fram Strait recirculation and Arctic Intermediate Waters (AIW), and occurs where baroclinic instability promotes enhanced mixing along the western perimeter of the Greenland Sea gyre; this conforms with the argument¹⁵ that an input of AIW from the convective gyres is needed to explain the observed oxygenation of the boundary current between Fram Strait and Denmark Strait. Finally, a detailed recent investigation¹⁶ concludes that DSOW is composed of equal

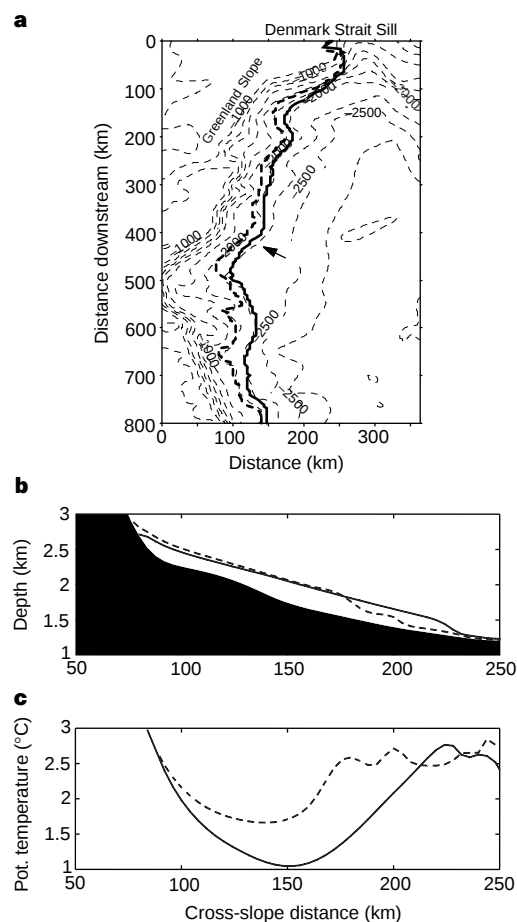


Figure 3 Simulating the downstream response of the overflow plume to an imposed 1°C warming at the sill. **a**, Location of the core of the Denmark Strait overflow south of the sill in relation to bathymetry in two runs of a numerical overflow model³. The cold run (solid lines) assumes temperature = -0.3°C, salinity = 34.90 at the sill, and in the warm run (dashed lines), sill temperatures are raised by 1°C to +0.7°C while salinities are held unchanged. The arrow marks the approximate range of the current-meter array off Angmagssalik. At this point, panels **b** and **c** show the cross-current distributions of layer thickness and potential temperature for the two runs. The upslope movement of the plume and the thinning of its lower edge are similar to changes observed in winter 1996–97.

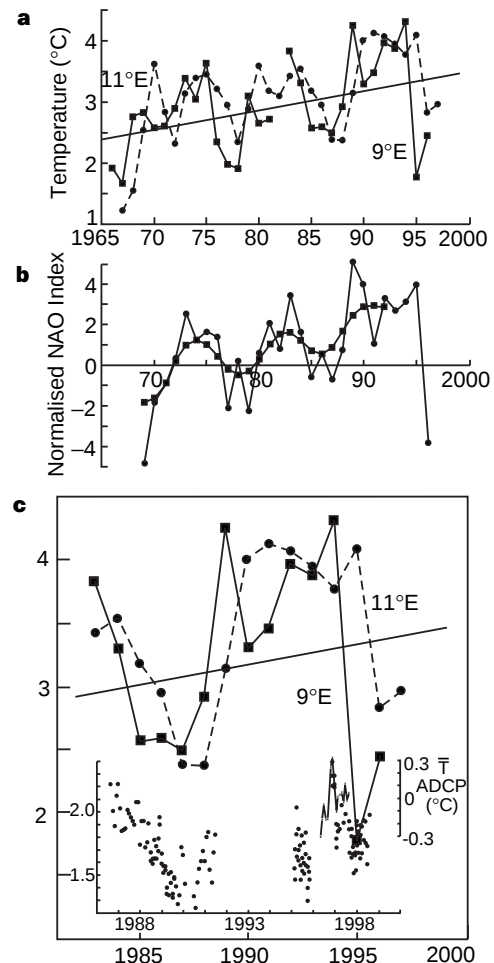


Figure 4 Evidence that variability in the Denmark Strait overflow may be linked to earlier climatic and hydrographic changes upstream. **a**, Mean 50–500 m temperature in August–September at 9° E and 11° E on a section across the West Spitzbergen Current in the eastern Fram Strait at 76° 20' N, compared with the normalized NAO Index⁵ (**b**). In **c**, West Spitzbergen Current temperatures since 1983 are compared in detail with the near-bottom temperature record from current meters set at 2000 m depth in the core of the Denmark Strait overflow off Angmagssalik, three years later. (Note that in 1990–91, part of the latter time series is taken from a more northerly array across the east Greenland continental slope (mooring no. 9009-2; ref. 2) and is hence from an instrument at slightly shallower water depths (1,700 m), but from the equivalent position in the core of the overflow, and after changes in water-mass properties due to entrainment are largely complete. The records are thus comparable.) The thin solid line shown is the thermistor record (°C) from a bottom-mounted Acoustic Doppler Current Profiler (ADCP) set on the sill of the Denmark Strait during 1996–7. Transit times between this sill and the deepening overflow off Angmagssalik are assumed to be brief so that this record and the current-meter thermistor records off Angmagssalik are comparable without any intervening time-lag.

thirds of NAW, AIW and Arctic Ocean Deep Water (AODW). Since there is evidence that the balance of vertical and horizontal exchange has altered within the Nordic seas^{17,18}, it is possible—even likely—that the DSOW recipe is not fixed or definable but is itself time-dependent. Nevertheless, the full range of modern DSOW recipes all contain a sufficient contribution of NAW to justify the idea that changes in DSOW characteristics may be linked to fluctuations in the NAW at some appropriate remove in time and space.

Second, though the available records are uncomfortably short and 'gappy', we suggest there is evidence that the ocean temperature series at either end of this circuit are correlated, once an appropriate advective–diffusive time-lag is interposed. In Fig. 4, we compare the 50–500 m mean temperature of the West Spitzbergen Current in Fram Strait (the northward extension of the Norwegian Atlantic Current) with a time-series of current meter thermistor records from the Denmark Strait overflow at 2,000 m depth off Angmagssalik, some 3 years later and some 2500 km further south. The main features of the two series—the long decline in temperature in the mid-to-late 1980s, the temperature maximum in the mid-to-late 1990s, and the recent rapid cooling associated with the brief recurrence of extreme negative NAO conditions in 1996—seem to correspond at each site.

The three-year advective–diffusive time-lag between the two sites remains to be tested. At present, we have little more basis for this than the fact that another significant ocean climate signal, the "Great Salinity Anomaly"¹⁹, took about this time to spread the equivalent distance from the Faroe–Shetland Channel (1976) to Fram Strait (1979). However, we may soon determine the rate of the Nordic seas circulation directly using two novel tracers: Sellafield ⁹⁹Tc spreading around the margins in the Norwegian Atlantic Current, and SF₆ purposely released into the convective centre of the Greenland Sea by the ESOP-II programme of EC-MAST. An experimental bottom-mounted whole-water sampler has been moored at 2,000 m depth in the core of the overflow to record their time of arrival.

Although the demonstration of some physical linkage between Fram Strait and Denmark Strait is not irrelevant to the prognosis, and the lagged correlation in temperature between the two sites is encouraging, these two results fall far short of demonstrating whether (and to what extent) the hydrographic character of DSOW may be set by events upstream (and hence be predictable). The most formidable objection is evident in Fig. 4—that however well-correlated their time-series may be, the transfer of an ocean-climate anomaly from Fram Strait to Denmark Strait must somehow survive the water-mass transformations implied by a change in mean temperature (from 3 °C in Fram Strait through 0 °C at the Denmark Strait sill, to 2 °C at depth off Angmagssalik). The modelling of water-mass mixing and exchange processes around this circuit is still insufficiently advanced to comment usefully on variability in either speed or hydrography. We note merely that in the model runs described above, about half of the imposed 1 °C temperature change survived the process of entrainment south of the Denmark Strait sill, which must represent the most intensive process of water-mass transformation on the circuit from Fram Strait. So for the present, the important aim of predicting overflow variability is likely to be led by observations rather than models. □

Received 2 July; accepted 16 October 1998.

- Dickson, R. R., Gmitrowicz, E. M. & Watson, A. J. Deep water renewal in the northern North Atlantic. *Nature* **344**, 848–850 (1990).
- Dickson, R. R. & Brown, J. The production of North Atlantic Deep Water: sources, rates and pathways. *J. Geophys. Res.* **99**, 12319–12341 (1994).
- Jungclaus, J. H. & Backhaus, J. O. Application of a transient, reduced-gravity, plume model to the Denmark Strait Overflow. *J. Geophys. Res.* **99**, 12375–12396 (1994).
- Cayan, D. R. Latent and sensible heat flux anomalies over the northern oceans: the connection to monthly atmospheric circulation. *J. Clim.* **5**, 354–369 (1992).
- Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
- Barnston, A. G. & Livezey, R. E. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Weath. Rev.* **115**, 1083–1126 (1987).

- Hurrell, J. W. & van Loon, H. Decadal variations in climate associated with the North Atlantic Oscillation. *Clim. Change* **36**, 301–326 (1997).
- Dickson, R. R. *et al.* The Arctic Ocean response to the North Atlantic Oscillation. *J. Clim.* (submitted).
- Swift, J. H. & Aagaard, K. Seasonal transitions and water mass formation in the Iceland and Greenland Seas. *Deep-Sea Res.* **28**, 1107–1129 (1981).
- Smethie, W. M. & Swift, J. H. The tritium: krypton-85 age of Denmark Strait Overflow Water and Gibbs Fracture Zone Water just south of Denmark Strait. *J. Geophys. Res.* **94**, 8265–8275 (1989).
- Swift, J. H. & Koltermann, K.-P. The origin of Norwegian Sea Deep Water. *J. Geophys. Res.* **93**, 3563–3569 (1988).
- Aagaard, K., Fahrbach, E., Meincke, J. & Swift, J. H. Saline outflow from the Arctic Ocean: its contribution to the deep waters of the Greenland, Norwegian and Icelandic seas. *J. Geophys. Res.* **96**, 20433–20441 (1991).
- Mauritzen, C. Production of dense overflow waters feeding the North Atlantic across the Greenland–Scotland Ridge: Part 1. Evidence for a revised circulation scheme. *Deep-Sea Res.* **43**, 769–806 (1996).
- Strass, V. H., Fahrbach, E., Schauer, U. & Sellmann, L. Formation of Denmark Strait Overflow Water by mixing in the East Greenland Current. *J. Geophys. Res.* **98**, 6907–6919 (1993).
- Swift, J. H. Dissolved Oxygen in the Arctic Ocean and Nordic Seas. *Proc. ACSYS Conf.* **2**, 259–261 (1997).
- Rudels, B., Quadfasel, D. & Friedrich, H. J. The Arctic Circumpolar Boundary Current. *Deep-Sea Res.* (submitted).
- Meincke, J., Rudels, B. & Friedrich, H. J. The Arctic Ocean–Nordic Seas thermohaline system. *Int. Council Explor. Sea J. Mar. Sci.* **54**, 283–299 (1997).
- Dickson, R. R., Lazier, J., Meincke, J., Rhines, P. & Swift, J. Long-term coordinated changes in the convective activity of the North Atlantic. *Prog. Oceanogr.* **38**, 241–295 (1996).
- Dickson, R. R., Meincke, J., Malmberg, S.-A. & Lee, A. J. The "Great Salinity Anomaly" in the Northern North Atlantic, 1968–82. *Prog. Oceanogr.* **20**, 103–151 (1988).

Acknowledgements. This work was funded in part by the EC MAST-III VEINS programme, and in part by the SIO/LDEO Consortium of the NOAA Global Change Program.

Correspondence and requests for materials should be addressed to R.D. (e-mail: r.r.dickson@cefes.co.uk).

Mesozoic subducted slabs under Siberia

Rob Van der Voo^{*,†}, Wim Spakman[†] & Harmen Bijwaard[†]

^{*} Department of Geological Sciences, the University of Michigan, Ann Arbor, Michigan 48109-1063, USA

[†] Vening Meinesz School of Geodynamics, Institute of Earth Sciences, Utrecht University, Budapestlaan 4, 3584 CD Utrecht, the Netherlands

Recent results from seismic tomography demonstrate that subducted oceanic lithosphere can be observed globally as slabs of relatively high seismic velocity in the upper as well as lower mantle^{1,2}. The Asian mantle is no exception, with high-velocity slabs being observed downwards from the west Pacific subduction zones under the Kurile Islands, Japan and farther south^{3–5}, as well as under Asia's ancient Tethyan margin. Here we present evidence for the presence of slab remnants of Jurassic age that were subducted when the Mongol–Okhotsk and Kular–Nera oceans closed between Siberia, the combined Mongolia–North China blocks and the Omolon block^{6–8}. We identify these proposed slab remnants in the lower mantle west of Lake Baikal down to depths of at least 2,500 km, where they join what has been interpreted as a 'graveyard' of subducted lithosphere at the bottom of the mantle. Our interpretation implies that slab remnants in the mantle can still be recognized some 150 million years or more after they have been subducted and that such structures may be useful in associating geodynamic to surface-tectonic processes.

Eurasia is, without doubt, the next supercontinent in the making, as continental elements such as Australia and even the Americas are steadily approaching it and seem destined to join it in the future. The growth of Asia started, however, a long time ago, when significant additions to its Siberian core began to occur through continent–continent collisions and accretion of smaller continental blocks⁶. The Mongolia and North and South China blocks, having amalgamated in the Permian–Triassic periods¹⁰ (or possibly as late as the Jurassic period¹¹) joined Siberia in Late Jurassic¹² or Early Cretaceous times⁸ (Fig. 1). We will call the former ocean between Siberia and Mongolia the Mongol–Okhotsk Ocean, but it has also been called the Khangai–Khankey Ocean⁶. The northeastern-most Siberian block (Omolon) accreted to the growing Asian continent in

the Late Jurassic, closing the Kular–Nera Ocean⁶, and causing the Verkhoyansk Mountains¹³ (Fig. 1).

Improvements in tomographic techniques in the past few years have allowed more detailed visualization of deep fossil slabs, such as those that appear to have resulted from subduction of the Pacific and Tethyan oceans^{1,14}. Tomography parametrizes the mantle either by spherical harmonic functions or by using local cell slowness functions. The sizes of these cells in the latter parametrization have steadily decreased from early values¹⁵ of about 10° to recent values^{1,14,16} of about 2°–0.6°. Improvements in the travel-time data set¹⁷, the use of a large number of composite rays, and variably sized cells that minimize the difference in cell hit count¹⁸ have now¹⁴ provided tomographic images with great detail. Examples of the resolution, data fit, and sensitivity tests for the tomographic results we interpret here are more fully described elsewhere¹⁴.

Figures 2 and 3 show map views at various lower-mantle depths

and a cross-section under Siberia. The Mongol–Okhotsk and Verkhoyansk sutures are marked by a P-wave velocity contrast in the crust and uppermost mantle (not shown), also visible in surface wave studies^{19,20}. Clear anomalies, well resolved both vertically and horizontally, are seen at depths greater than 1,200 km (Fig. 2). In the crust, the Mongol–Okhotsk–Verkhoyansk suture runs from near the northern shore of Lake Baikal to the Sea of Okhotsk (for locations, see Fig. 2), and from there northwards under the Verkhoyansk Mountains. At depths greater than 1,500 km, the velocity anomalies are pronounced; amplitude variance tests indicate that the variation (of about +0.3% or higher) is well above amplitude noise level in the lower mantle¹⁴. At 1,500 km depth, the anomalies form a ‘hook’, running first west–northwestwards from Mongolia and then northwards towards the Siberian Arctic coast. At depths of 1,900–2,300 km, the anomalies are gradually more displaced to the west, and display an overall ‘Z’-shaped feature at 2,300 km (see curved yellow line in Fig. 2). At 2,500 km and below, the high-velocity anomalies merge with a broad east Asian high-velocity area in the lowermost mantle (Figs 2 and 3), which has been called a ‘graveyard’ of slabs⁹, an interpretation corroborated by our results.

In the cross-section (Fig. 3, top), this feature (M) is illustrated side-by-side with the currently active subduction zone (P) under northern Japan. Below 1,400 km, we see the southern leg of the hook to the northwest of Mongolia portrayed all the way down to the core–mantle boundary. Figure 3 also illustrates, for this same cross-section, how a simulated input of a layer-cake model¹⁴ is resolved in our inversion. The figure shows that below depths of 1,500 km, the M and P anomalies are both well resolved. The overall amplitude and spatial characteristics of the high P-wave velocity anomalies in the deep mantle under Siberia are comparable to those observed around the Pacific and under the Mediterranean–Himalaya–Indonesian belt and warrant a similar interpretation: as remnants of subducted slabs. Some positive anomalies are imaged between 400 and 1,200 km depth, but these are spatially scattered, of low amplitude, and not resolved. For instance, the ‘connection’ between upper and mid-mantle (>1,200 km) anomalies in Fig. 3 (top) is rather local and unresolved.

Identification of the Pacific and Tethyan subduction zones in the maps and section is rather straightforward^{1,2,14}; thus, we are confident that the high-velocity zone we have described under Siberia is not related to Cenozoic subduction. However, elimination of other, Mesozoic, subduction zones under Asia as candidates for the deep-mantle portions of the Siberian slab is more difficult: this is because the Asian lithosphere may have drifted significantly with respect to the deeper mantle formerly underlying it. Because of the increasing certainty that the hotspots have moved significantly with respect to each other, they do not provide a suitable framework to assess this issue^{21,22}. However, the western and eastern Pacific subduction zones and their deep slabs provide an alternative approach. The western Pacific subducted slab dips westwards and then sinks down vertically in the lower mantle, without large apparent displacements from where one would expect to find it (Figs 2, 3). In contrast, the deep-mantle slabs resulting from east Pacific subduction under North America in late Mesozoic–Palaeogene times can be recognized under the Great Lakes area and east of there, as far away as New England^{21,14}. The North American lithosphere has clearly moved far to the west with respect to these deep-mantle slabs. From (1) these relative displacements of the two slabs with respect to the present lithospheric locations where they started their downward journeys, and (2) the displacements of the continents during the opening of the Atlantic Ocean, we can deduce that the Siberian lithosphere has moved very little in an east–west direction with respect to the deep-mantle slabs; we can also deduce that the Atlantic Ocean has spread primarily by westward drift of North America, which has overridden the slabs under it to the extent of some 3,500 km or more. Moreover, other deep slab remnants under

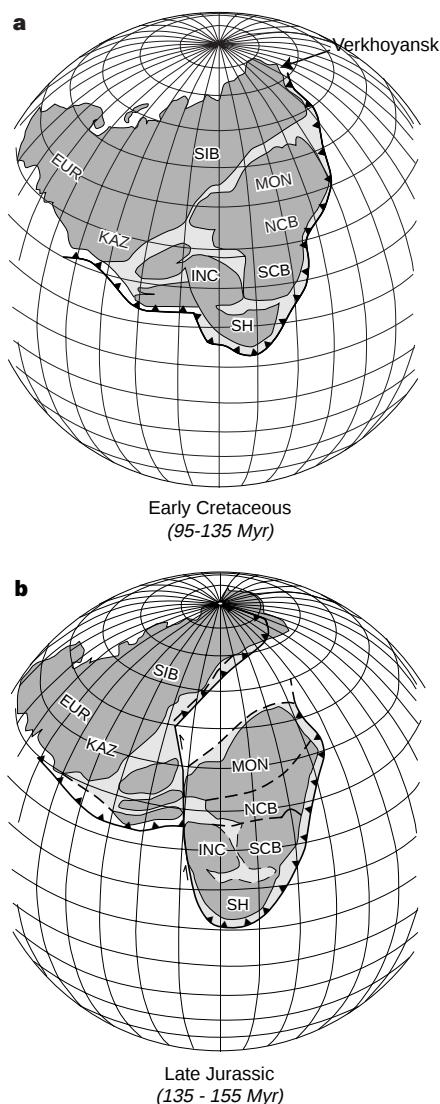


Figure 1 Early Cretaceous and Late Jurassic palaeoreconstructions of Siberia (SIB) and adjacent Asian blocks⁸. (Redrawn from ref. 8 with permission from the authors, R. Enkin *et al.*) **a**, Early Cretaceous; **b**, Late Jurassic. Late Jurassic subduction of the Mongol–Okhotsk Ocean is shown as occurring under the southeastern and northeastern (Verkhoyansk) margins of Siberia. Subduction probably also occurred under the Mongolian margin⁶, but this is not shown here. Abbreviations of block names are as follows: EUR, Europe; KAZ, Kazakhstan; MON, Mongolia; NCB and SCB, North and South China blocks; INC, Indochina; and SH, Shan–Thai block⁸.

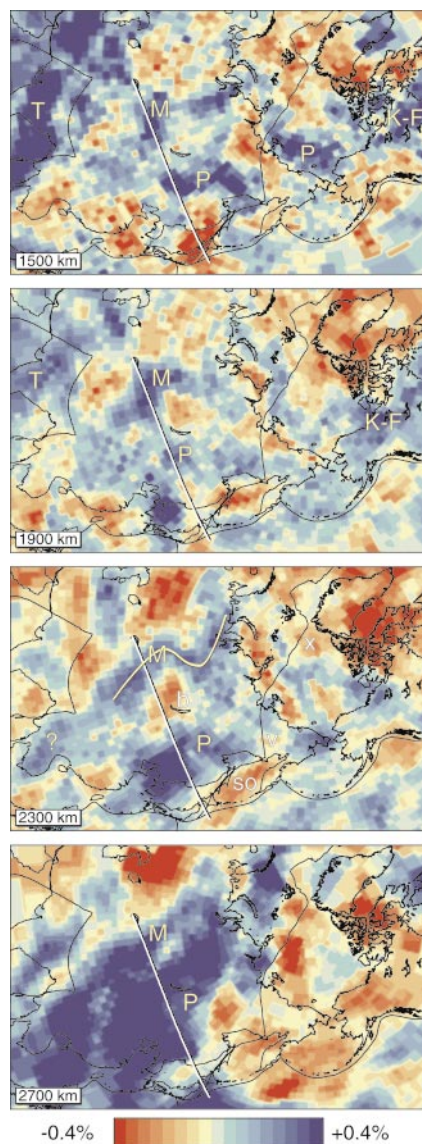


Figure 2 Tomographic P-wave velocity anomaly patterns in the deep mantle under Asia, for four different depths between 1,500 and 2,700 km. The superposed straight line indicates the location of the cross-section of Fig. 3. Velocity anomalies are displayed in percentages with respect to the average P-wave velocity at depth from model ak135²⁸. Capital letters represent our interpretation of the Mongol–Okhotsk (M), Pacific (P), Tethyan (T) and Kula–Farallon (K–F) oceanic lithospheric slabs in all four frames; lower-case letters in the 2,300 km frame represent geographical features: b, Lake Baikal; so, Sea of Okhotsk; v, Verkhoyansk suture; x, present-day North Pole. The question mark denotes a ‘slab’ of uncertain origin.

Eurasia due to Tethyan subduction are not shifted in an east–west sense. It seems justifiable, therefore, to infer that the deep-mantle slab we have described under Siberia is not related to subduction other than that of the Mongol–Okhotsk and Kular–Nera oceans.

In Fig. 4 we show the migration of this slab as a function of depth, and compare it with the palaeolatitudinal location of the lithospheric Mongol–Okhotsk–Verkhoyansk suture zone as a function of geological time. This figure shows that the orientations of the present-day suture zones at the surface parallel a higher-velocity contrast within the lithosphere down to depths of at least 95 km. At depths of 1,200 km or more, the velocity anomalies are rotated anticlockwise from the surface orientation. However, their locations correspond to the (palaeogeographically restored) position of the eastern and southern margin of Siberia in early Tertiary and older

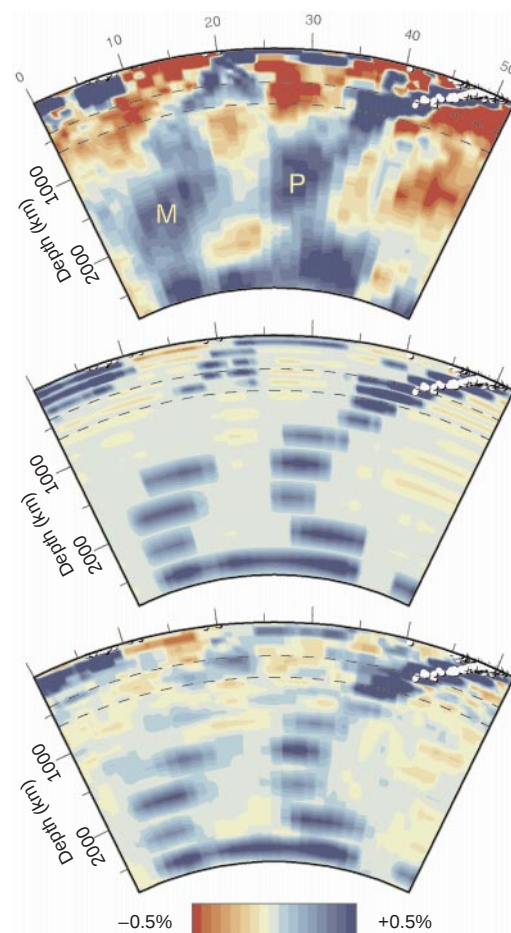


Figure 3 Cross-section through tomographic model. Top, tomographic results displayed in a vertical section along the line indicated in Fig. 2, showing the Pacific (P) subduction zone under Japan (right) and the inferred Mongol–Okhotsk slab (M) discussed here (left) side-by-side. White dots show locations of earthquakes in the west Pacific Benioff zone. Middle, a simulated input of a whole-mantle layer-cake model¹⁴ for this same cross-section, which is resolved and (partially to completely) reproduced in our inversion analysis (bottom).

times, but are further to the west, presumably owing to the angle of generally westward subduction during Late Jurassic times (Fig. 4).

The ‘Z’-shape of the slabs below 2,000 km (Fig. 2) is interpreted as the result of a similarly ‘Z’-shaped surficial subduction pattern of the Kular–Nera and Mongol–Okhotsk oceans not only under Siberia itself⁶ (Fig. 1), but under Siberia’s Jurassic northeastern and southeastern margins, as well as under the northern margin of Mongolia (see Fig. 21.39 in ref. 6).

Given that subduction ceased about 150 Myr ago, the depth of about 1,500 km for the top of the deep slab under Siberia implies an averaged sinking velocity of 1 cm yr^{-1} since the Jurassic period, which is not very different from sinking velocities inferred in some previous studies², although higher rates of $\sim 3 \text{ cm yr}^{-1}$ have also been proposed²³. Subduction velocities are thought to reduce by a factor of about four when slabs begin to penetrate into the higher-viscosity deeper mantle²⁴; a sinking velocity of 1 cm yr^{-1} would then imply a surficial convergence rate of 4 cm yr^{-1} . Even if subduction occurred on only one (for example, the Siberian) side of the Mongol–Okhotsk Ocean, the width of this ocean may have been at least 2,000 km at 200 Myr, which matches plate reconstructions²⁵.

We argue that the high-velocity structures under Siberia are logically interpreted as remnants of oceanic lithosphere that subducted before the Early Cretaceous and that, therefore, subducted lithosphere of Jurassic age can still be recognized after penetration

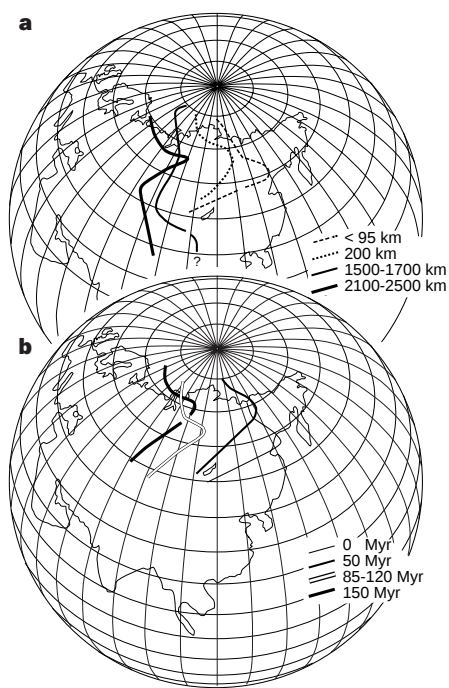


Figure 4 A comparison of the locations of tomographic velocity anomalies and the expected palaeolocations of the Siberian active margins. **a**, Locations of the lithospheric suture and the main axis of the ocean-lithospheric slab remnants as a function of depth, as determined from our tomographic results. **b**, Present-day and palaeogeographically reconstructed locations of the Mongol-Okhotsk-Verkhoyansk suture zones as a function of time (longitudes are arbitrary), using the palaeomagnetic pole determinations for Siberia of Zhao *et al.*²⁹ (81°N, 158.6°E for 50 Myr; 73.8°N, 202.4°E for 85–120 Myr; 70.1°N, 184.3°E for 150 Myr).

into the lower mantle even after subduction stopped some 150 Myr ago. Whether this visibility is a result of temperature²⁶, composition, pressure, or a combination thereof, is an open question, but what is clear is that the Mongol–Okhotsk subducted lithospheric material has been feeding the ‘graveyard’ of slabs under Asia. Our conclusions also imply that significant downwelling is a characteristic of growing supercontinents for hundred of millions of years²⁷, and that most, if not all, significantly fast anomalies in the deeper mantle appear to be associated with past subduction. This renders tomography an important tool for testing palaeogeographical reconstructions. □

Received 23 June; accepted 27 October 1998.

- van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. Evidence for deep-mantle circulation from global tomography. *Nature* **386**, 578–584 (1997).
- Grand, S., van der Hilst, R. D. & Widiyantoro, S. Global seismic tomography: A snapshot of convection in the Earth. *GSA today* **7**, 1–7 (1997).
- van der Hilst, R., Engdahl, R., Spakman, W. & Nolet, G. Tomographic imaging of subducted lithosphere below the northwest Pacific island arcs. *Nature* **353**, 37–43 (1991).
- Richards, M. A. & Engebretson, D. C. Large-scale mantle convection and the history of subduction. *Nature* **355**, 437–440 (1992).
- Wen, L. & Anderson, D. L. The fate of slabs inferred from seismic tomography and 130 million years of subduction. *Earth Planet. Sci. Lett.* **133**, 185–198 (1995).
- Şengör, A. M. C. & Natal'in, B. A. in *The Tectonic Evolution of Asia* (eds Yin, A. & Harrison, M.) 486–640 (Cambridge Univ. Press, 1996).
- Zhao, X. X., Coe, R. S., Zhou, Y., Wu, H. & Wang, J. New paleomagnetic results from northern China: Collision and suturing with Siberia and Kazakhstan. *Tectonophysics* **181**, 43–81 (1990).
- Enkin, R., Yang, Z. Y., Chen, Y. & Courtillot, V. Paleomagnetic constraints on the geodynamic history of the major blocks of China from the Permian to the Present. *J. Geophys. Res.* **97**, 13953–13989 (1992).
- Wyssession, M. E. in *Subduction* (eds Bebout, G. *et al.*) 369–384 (Monogr. 96, Am. Geophys. Union, Washington DC, 1996).
- Yin, A. & Nie, S. Y. in *The Tectonic Evolution of Asia* (eds Yin, A. & Harrison, M.) 442–485 (Cambridge Univ. Press, 1996).
- Gilder, S. & Courtillot, V. Timing of the North-South China collision from new middle to late Mesozoic paleomagnetic data from the North China block. *J. Geophys. Res.* **102**, 17713–17727 (1997).
- Ziegler, A. M. *et al.* in *The Tectonic Evolution of Asia* (eds Yin, A. & Harrison, M.) 371–400 (Cambridge Univ. Press, 1996).
- Parfenov, L. M. Tectonics of the Verkhoyansk-Kolyma Mesozoids in the context of plate tectonics. *Tectonophysics* **199**, 319–342 (1991).

- Bijwaard, H., Spakman, W. & Engdahl, R. Closing the gap between regional and global travel time tomography. *J. Geophys. Res.* (in the press).
- Dziewonski, A. M., Hager, B. H. & O’Connell, R. J. Large-scale heterogeneities in the lower mantle. *J. Geophys. Res.* **82**, 239–255 (1977).
- Zhou, H. A high resolution P-wave model for the top 1200 km of the mantle. *J. Geophys. Res.* **101**, 27791–27810 (1996).
- Engdahl, E. R., van der Hilst, R. D. & Buland, R. P. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull. Seismol. Soc. Am.* **88**, 722–743 (1998).
- Spakman, W. & Bijwaard, H. Irregular cell parameterization of tomographic problems. *Ann. Geophys.* **16** (suppl. 1), 28 (1998).
- Curtis, C., Dost, B., Trampert, J. & Snieder, R. Eurasian fundamental mode surface wave phase velocities and their relationship with tectonic structures. *J. Geophys. Res.* (in the press).
- Ritzwoller, M. H. & Levshin, A. L. Eurasian surface wave tomography: Group velocities. *J. Geophys. Res.* **103**, 4839–4878 (1998).
- Cande, S. C., Raymond, C. A., Stock, J. & Haxby, W. F. Geophysics of the Pitman Fracture Zone and Pacific–Antarctic plate motions during the Cenozoic. *Science* **270**, 947–953 (1995).
- Tarduno, J. A. & Cottrell, R. D. Paleomagnetic evidence for motion of the Hawaiian hotspot during formation of the Emperor Seamounts. *Earth Planet. Sci. Lett.* **153**, 171–180 (1997).
- Ricard, Y., Richards, M. A., Lithgow-Bertelloni, C. & Le Stunff, Y. A geodynamic model of mantle density heterogeneity. *J. Geophys. Res.* **98**, 21895–21909 (1993).
- Lithgow-Bertelloni, C. & Richards, M. A. The dynamics of Cenozoic and Mesozoic plate motions. *Rev. Geophysics* **36**, 27–78 (1998).
- Scotese, C. R. & Golonka, J. *Paleomap Paleogeographic Atlas* (Paleomap project, Univ. Texas, Arlington, 1992).
- Zhong, S. & Gurnis, M. Mantle convection with plates and mobile, faulted plate margins. *Science* **267**, 838–843 (1995).
- Gurnis, M. Large-scale mantle convection and the aggregation and dispersal of supercontinents. *Nature* **232**, 695–699 (1988).
- Kennett, B. L. N., Engdahl, E. R. & Buland, R. Constraints on seismic velocities in the Earth from travel times. *Geophys. J. Int.* **122**, 108–124 (1995).
- Zhao, X. X., Coe, R. S., Gilder, S. A. & Frost, G. M. Palaeomagnetic constraints on palaeogeography of China: Implications for Gondwanaland. *Austr. J. Earth Sci.* **43**, 643–672 (1996).

Acknowledgements. We thank M. Richards and M. Gurnis for constructive comments. H.B. was supported by the Netherlands Organization for Scientific Research (NWO).

Correspondence and requests for materials should be addressed to R.V.d.V. (e-mail: voo@umich.edu).

Can aposematic signals evolve by gradual change?

Leena Lindström, Rauno V. Alatalo, Johanna Mappes, Marianna Riipi & Laura Vertainen

University of Jyväskylä, Department of Biological and Environmental Science, Konnevesi Research Station, PO Box 35, FIN-40 351 Jyväskylä, Finland

Aposematic species, which signal conspicuously of their unprofitability to predators, have puzzled evolutionary biologists for over a century^{1,2}. Although conspicuousness of unpalatable prey improves avoidance learning by predators^{3–5}, it also involves an evolutionary paradox: with increasing detectability^{4,6–8} the deviant aposematic prey would suffer high predation initially from naive predators. Here we test a neglected idea^{7–11} that aposematic coloration may evolve by gradual change rather than by major mutations. Weak signals did not suffer high initial predation, but predators (great tits, *Parus major*) did not learn to separate them from cryptic palatable prey. Furthermore, enhanced avoidance of more conspicuous signals occurred only if predators had previously encountered relatively strong signals. Thus, the gradual-change hypothesis does not provide an easy solution to the initial evolution of aposematism through predator learning. However, the possibility remains that cost-free step-wise mutations over the range of weak signals could accumulate under neutral selection to produce effective strong signals.

It has been assumed that sudden, pronounced mutations turned unpalatable cryptic prey into highly conspicuous forms. This hypothesis of sudden change may account for why grouping has been suggested to be essential in explaining how brightly coloured insects first evolved^{12–14}. Naive predators encountering a group are likely to leave some of the unpalatable prey untouched, whereas any single solitary individual may not escape being predated. Alternatively, gradual change in prey detectability^{7–9,11}, initially not changing the cryptic type into an overtly conspicuous prey type, may allow evolution of aposematism even in solitary prey. This

Table 1 Average detectability scores

Signal type*	Visibility experiment			Learning experiment			Test	
	$\bar{x}$	s.e.m.	n	$\bar{x}$	s.e.m.	n	U	P
✕	1.16	0.15	12	1.14	0.15	8	44	0.792
✕	1.56	0.10	12	0.88	0.19	7	9	0.015
■	2.09	0.08	12	1.26	0.20	8	10	0.009
■	2.72	0.22	12	2.16	0.36	8	33	0.270

* The four symbols represent a detectability continuum from the weakest signal type (first row) to the most conspicuous signal type (last row). Average detectability scores reflect the visibility risk for the non-cryptic prey types compared with the cryptic type. Detectability scores for the visibility experiment are averages from two trials. Because unpalatability might interfere with the visibility risks, the detectability scores for the initial phase of the learning experiment are counted from the five first prey items taken on the first trial. We compared these two situations (Mann-Whitney *U*-tests, Bonferroni corrected *P*-values) for each symbol type. There seemed to be a slight reduction in visibility risks during the learning experiment which might result from neophobia.

hypothesis was modelled recently by Yachi and Higashi¹⁰.

Using the novel-world method¹³, where naive predators are presented with warning signals not found in their environment, we conducted three experiments to test the gradual-change hypothesis. We studied the visibility differences, selective advantages of learning and generalization of avoidance of different levels of conspicuousness by creating a detectability continuum of artificial prey types. We created one cryptic signal and four aposematic signals, which ranged from nearly cryptic to highly conspicuous in relation to their background (see Fig. 2 for signals). Wild great tits (*Parus major*) were used as predators. To test the visibility differences of the five signals, eight palatable solitary items of each signal type were presented simultaneously. Birds received all signal types before the test in order to avoid any neophobic reactions. There were clear visibility differences between the signal types ($n = 12$, Friedman $\chi^2 = 25.13$, d.f. = 4, $P < 0.001$; Table 1), except for the cryptic type and the weakest aposematic signal type (after sequential Bonferroni corrections; Wilcoxon $T = -0.82$, $P = 0.500$). The most conspicuous signal type was observed by the birds nearly three times more often than the cryptic type, which highlights the increased detectability risk.

A central question in understanding the evolution of warning coloration is the capability of predators to distinguish unpalatable from palatable prey. To test this, we devised a learning experiment in which a new set of birds was divided into four treatments according to the aposematic signal type. Before the experiment, each bird was offered three cryptic prey items to mimic the situation that only the aposematic prey signal was unfamiliar. The birds were then tested on five consecutive days for their ability to learn to avoid their assigned signalling unpalatable prey from the palatable cryptic prey. The initial detectability differences were not as clear as in the visibility experiment where birds had experienced all prey types

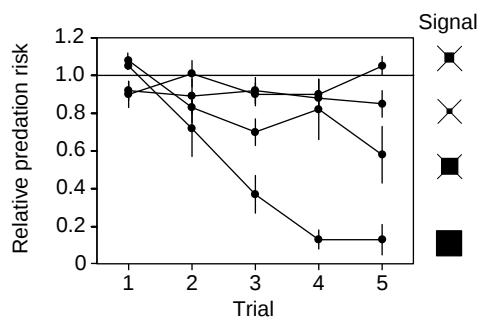


Figure 1 The mean relative predation risk for all four aposematic signal types on five consecutive days in the learning experiment. The smallest square at the centre of cross is the weakest signal type, whereas the square that completely covers the cross is the most conspicuous signal type. The reference line indicates equal predation of cryptic and aposematic type. The bars represent standard errors of the mean (s.e.m.).

before testing (Table 1). The slight reduction in detectability differences might originate from neophobia^{15,16}. There was also a significant interaction between the amount of unpalatable prey that birds ate in the five consecutive days in different signal treatments (repeated measure analysis of variance: learning $\times$ signal, $F(8, 36) = 2.74$, $P = 0.018$). Although the least conspicuous aposematic signals do not suffer increased detectability risk, predators do not learn to avoid them (Fig. 1). The most detectable signals suffered higher predation at first, but they gained from avoidance learning. Learning was only significant when the signals were highly conspicuous (after Bonferroni corrections: second-most conspicuous signal, Friedman $\chi^2 = 14.16$, d.f. = 4, $P = 0.027$; most conspicuous signal, Friedman $\chi^2 = 19.81$, d.f. = 4, $P = 0.002$) and thus it seems that only the strongest signals can be avoided effectively.

One possible mechanism for the gradual-change pathway to evolve is through avoidance of even stronger signals by a process similar to a peak-shift effect^{17,18}. Under this hypothesis, predators first learn to avoid a given signal and subsequently avoid exaggerated new signals even more. To test this idea, the birds from the learning experiment were presented with all four aposematic prey types simultaneously along with the cryptic palatable type from the cryptic background.

Throughout the learning experiment and training, the slight difference between the cryptic prey and the weakest signal was not sufficient for avoidance learning. The avoidance of more conspicuous signals than the weakest signal was not enhanced by any generalization effect and therefore the gradual-change route does not seem likely in the learning scenario (Friedman $\chi^2 = 6.91$,

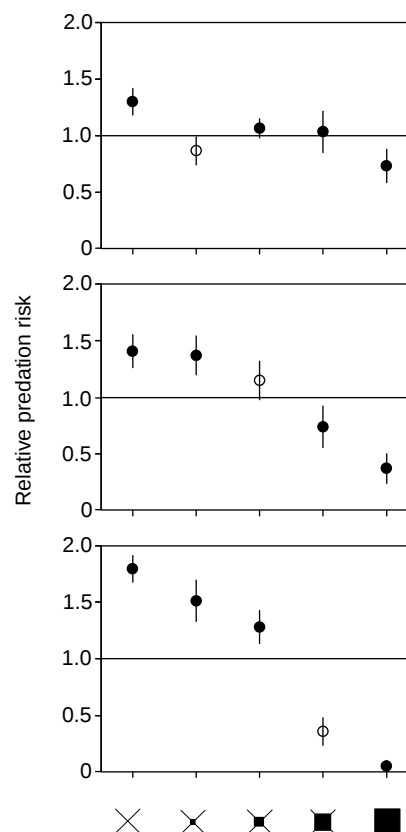


Figure 2 The mean relative predation risk of different signals in the generalization experiment. The symbols along the abscissa are signal types. They represent a detectability continuum from the cryptic type (far left) through the four aposematic signal types from the weakest signal to the most conspicuous signal type (far right). The open circles indicate the signal that the birds had been trained to encounter as unpalatable. The bars represent s.e.m.

d.f. = 4, $P = 0.141$; Fig. 2). Although the next signal (the third-strongest signal) was clearly detectable in the visibility experiment, birds had been unable to associate such prey as unpalatable during the learning experiment. After the additional training, birds shifted the avoidance of the learnt signal for the more conspicuous prey type (Friedman $\chi^2 = 15.34$, d.f. = 4, $P = 0.004$). In the continuum, the first signal that birds learnt to avoid (the second-strongest signal) also had the advantage in the generalization experiment and this avoidance benefited the most conspicuous signal (Friedman $\chi^2 = 37.5$, d.f. = 4, $P = 0.001$).

Weak signals that were not learnt did not suffer from increased predation, and thus any such mutations would not have been selected against. Thus, it is possible that stepwise mutations could occur by random drift until a level of conspicuousness is reached where predators can learn the signal. After this critical step the enhanced avoidance of predators might be generalized to more conspicuous signals without any initial detectability cost for the prey. □

Methods

The novel-world method allowed us to use wild great tits (*Parus major*) as naive predators. Birds were mist-netted (Jan–May 1997) near the Konnevesi research station and housed individually. Birds were trained to eat artificial prey items, pieces of almond placed between two 10 mm × 10 mm pieces of white paper that were glued together. In the experiments the prey items had printed symbols. All birds were familiarized to the testing situation by being allowed to feed on sunflower seeds and peanuts in the aviary before the experiment. A novel landscape was created in a 2.5 m × 5.2 m aviary, which was covered by white paper printed with 8 mm × 8 mm black crosses (580 crosses per m²). The cryptic prey items were made more difficult to detect by gluing fake prey items with cryptic symbols on the landscape (80 fakes per m²).

Visibility experiment. The continuum of signals from cryptic to conspicuous types was presented simultaneously on the novel landscape. All signals were palatable and eight items of each were used. Birds had previously received all of the different prey types (ten of each) in their housing cages to avoid neophobic reactions. The detectability scores were calculated from the first 20 prey items taken. The scores were calculated as a cumulative sum; the preferred prey type was given a value of 20 and the second-preferred prey type 19, and so on. The relations of the sums reflect the relative detectability risks of each prey type.

Learning experiment. Birds were tested individually on five consecutive days. Only one aposematic symbol was used over the course of the experiment for each bird. Cryptic prey items were palatable and aposematic prey items were made unpalatable by soaking almonds in a 40% solution of chloroquine. Prey items (24 aposematic and 24 cryptic) were scattered randomly on the landscape: the arena was divided into six equal-sized blocks, and four aposematic and four cryptic prey items were placed randomly in each block. Birds were allowed to eat 15 prey items in each daily experimental trial.

Generalization experiment. Avoidance towards the same unpalatable prey types that each bird had experienced in the learning experiment, was enhanced by offering further sets of cryptic and unpalatable prey (four replicates of five aposematic and five cryptic prey) simultaneously in their housing cages. In the experiment all aposematic prey types were presented simultaneously with cryptic items (eight items of each type). Birds were allowed to eat 15 prey items. After the experiments the birds were released back to sites where they were captured from.

Received 14 August; accepted 9 November 1998.

- Wallace, A. R. *Trans. Entomol. Soc. Lond.* 3rd Ser. 3, lxxx–lxxxi (1867).
- Schuler, W. & Roper, T. J. Responses to warning coloration in avian predators. *Adv. Study Behav.* 21, 111–146 (1992).
- Gittleman, J. L. & Harvey, P. H. Why are distasteful prey not cryptic? *Nature* 286, 149–150 (1980).
- Guilford, T. In *Insect Defences* (eds Evans, D. L. & Schmidt, J. O.) 23–61 (State Univ. New York Press, New York, 1990).
- Guilford, T. & Dawkins, M. Receiver psychology and the evolution of animal signals. *Anim. Behav.* 41, 1–14 (1991).
- Turner, J. R. G. & Mallet, J. L. B. Did forest islands drive the diversity of warningly coloured butterflies? Biotic drift and the shifting balance. *Phil. Trans. R. Soc. Lond. B* 351, 835–845 (1996).
- Endler, J. A. Frequency-dependent predation, crypsis and aposematic coloration. *Phil. Trans. R. Soc. Lond. B* 319, 505–523 (1988).
- Endler, J. A. In *Behavioural Ecology* (eds Krebs, J. R. & Davies, N. B.) 163–196 (Blackwell Scientific, Oxford, 1993).

- Mallet, J. & Singer, M. C. Individual selection, kin selection, and the shifting balance in the evolution of warning colours: the evidence from butterflies. *Biol. J. Linn. Soc.* 32, 337–350 (1987).
- Yachi, S. & Higashi, M. The evolution of warning signals. *Nature* 394, 882–884 (1998).
- Leimar, O., Enquist, M. & Sillén-Tullberg, B. Evolutionary stability of aposematic coloration and prey unprofitability: a theoretical analysis. *Am. Nat.* 128, 469–490 (1986).
- Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
- Alatalo, R. V. & Mappes, J. Tracking the evolution of warning signals. *Nature* 382, 708–710 (1996).
- Guilford, T. Is kin selection involved in the evolution of warning coloration? *Oikos* 45, 31–36 (1985).
- Coppinger, R. P. The effect of experience and novelty on avian feeding behavior with reference to the evolution of warning coloration in butterflies I. Reactions of wild-caught adult blue jays to novel insects. *Behaviour* 35, 45–60 (1969).
- Coppinger, R. P. The effect of experience and novelty on avian feeding behavior with reference to the evolution of warning coloration in butterflies II. Reactions of naive birds to novel insects. *Am. Nat.* 104, 323–335 (1970).
- Hanson, H. M. Effects of discrimination training on stimulus generalization. *J. Exp. Psychol.* 58, 321–334 (1959).
- Leimar, O. & Tuomi, J. Synergistic selection and graded traits. *Evol. Ecol.* 12, 59–71 (1998).

Acknowledgements. We thank C. Rowe and T. Guilford for insightful discussions; J. A. Endler, V. Kaitala, E. Koskela, K. Kärkkäinen, A. Lyytinen, T. Mappes and S. Parri for help in clarifying the manuscript; and the Konnevesi Research Station, Helinä Nisu and 'the academic hobby crafts club' for immense help. The study was supported by the Academy of Finland. Authors after senior author are in alphabetical order.

Correspondence and requests for materials should be addressed to L. Lindström (e-mail: lilema@tukki.jyu.fi).

The K⁺/Cl[−] co-transporter KCC2 renders GABA hyperpolarizing during neuronal maturation

Claudio Rivera^{*†}, Juha Voipio^{*}, John A. Payne[‡], Eva Ruusuvuori^{*}, Hannele Lahtinen^{*}, Karri Lamsa^{*}, Ulla Pirvola[†], Mart Saarma[†] & Kai Kaila^{*}

^{*} Department of Biosciences, Division of Animal Physiology, and [†] Institute of Biotechnology, Viikki Biocenter, University of Helsinki, FIN-00014, Helsinki, Finland

[‡] Department of Human Physiology, University of California School of Medicine, Davis, California 95616, USA

GABA (γ-aminobutyric acid) is the main inhibitory transmitter in the adult brain, and it exerts its fast hyperpolarizing effect through activation of anion (predominantly Cl[−])-permeant GABA_A receptors¹. However, during early neuronal development, GABA_A-receptor-mediated responses are often depolarizing^{2,3}, which may be a key factor in the control of several Ca²⁺-dependent developmental phenomena, including neuronal proliferation, migration and targeting^{4–6}. To date, however, the molecular mechanism underlying this shift in neuronal electrophysiological phenotype is unknown. Here we show that, in pyramidal neurons of the rat hippocampus, the ontogenetic change in GABA_A-mediated responses from depolarizing to hyperpolarizing is coupled to a developmental induction of the expression of the neuronal Cl[−]-extruding K⁺/Cl[−] co-transporter, KCC2 (ref. 7). Antisense oligonucleotide inhibition of KCC2 expression produces a marked positive shift in the reversal potential of GABA_A responses in functionally mature hippocampal pyramidal neurons. These data support the conclusion that KCC2 is the main Cl[−] extruder to promote fast hyperpolarizing postsynaptic inhibition in the brain.

In the rat hippocampus, fast GABAergic transmission is depolarizing at birth, only becoming hyperpolarizing and strictly inhibitory by the end of the first postnatal week^{2,3}. The molecular basis of this qualitative shift in GABA_A action has not been established.

Northern blots of total messenger RNA extracted from rat hippocampus showed a pronounced developmental upregulation of the expression of KCC2 mRNA (Fig. 1a, b). At postnatal day 0 (P0; day of birth) KCC2 mRNA (5.5 kilobases (kb)) was barely detectable. However, a steep increase in expression was evident at P5, reaching a level at P9 that was essentially similar to that in the

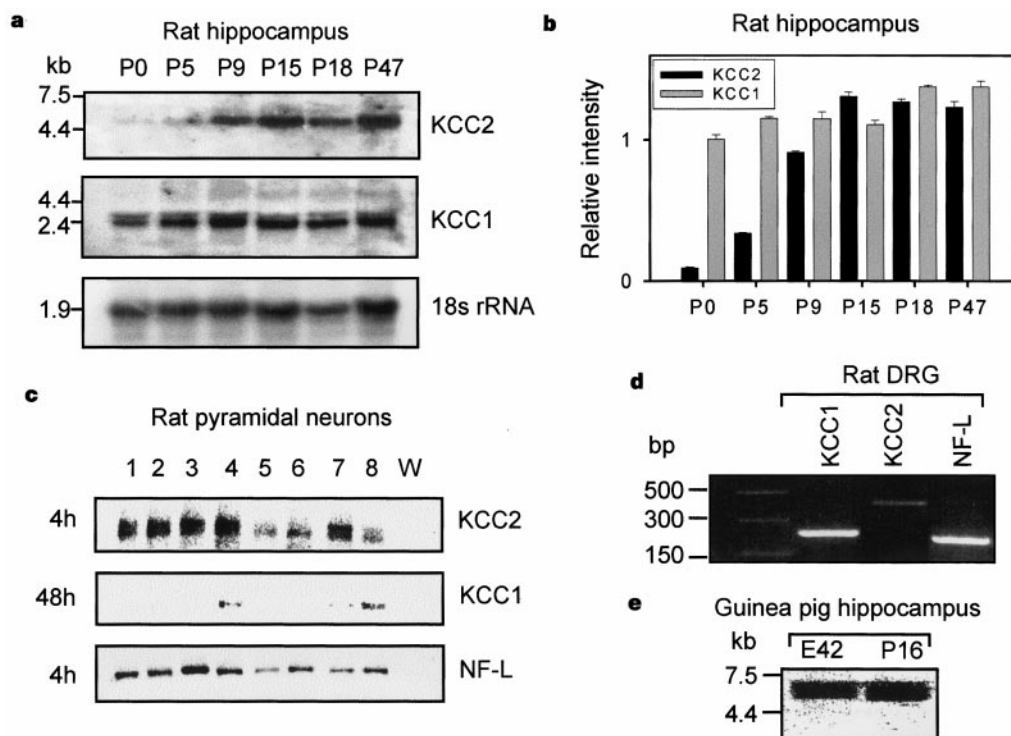


Figure 1 Developmental regulation and neuron-specific expression of the K^+/Cl^- co-transporter KCC2. **a**, Northern blot analysis of total mRNA isolated from rat hippocampus at different postnatal days shows a pronounced upregulation of KCC2 (but not KCC1) mRNA with a steep increase at around P5. 18s rRNA was used as control. **b**, Summary data from three northern blot experiments on rat hippocampus. The integrated intensities obtained by phosphorimaging were normalized to the values obtained from 18s rRNA. **c**, High expression of KCC2

compared with KCC1 mRNA in eight (lines 1–8) single rat hippocampal pyramidal neurons shown by Southern blot analysis of single-cell RT-PCR. KCC2 and NF-L mRNA blots were exposed for 4 h and KCC1 blots for 48 h (W, water control). **d**, RT-PCR analysis of KCC1, KCC2 and NF-L mRNA from P47 rat DRG. **e**, Northern blot analysis of KCC2 mRNA in the guinea-pig hippocampus at E42 and P16; note the rather stable level of expression.

adult (P47). In contrast, expression of the two KCC1 mRNAs (~2.4 kb and ~3.8 kb) was high already at P0, and did not change significantly during postnatal development.

The KCC1 isoform appears to be expressed ubiquitously in mammalian tissues⁸, whereas KCC2 is not found outside the brain⁷. Thus, it is possible that one, or both, of the two isoforms is expressed in pyramidal neurons. We examined this question in P16–P21 rat hippocampal pyramidal neurons using single-neuron reverse transcription with polymerase chain reaction (RT-PCR) followed by Southern blot analysis (Fig. 1c). Light neurofilament (NF-L) mRNA was used as a neuronal marker. Most (14 of 16 cells) of NF-L-positive pyramidal-like cells showed a prominent band for KCC2 with no detectable levels of KCC1 transcripts. Two neurons (shown in Fig. 1c) co-expressed KCC1 mRNA; this was detectable only after overexposing the films for 48 h. Expression of the KCC1 transcript alone was not observed in any of the NF-L-positive cells. To our knowledge, these results provide the first direct evidence for the postulated^{7,9} neuronal localization of KCC2, and they indicate that the KCC2 isoform is the main K^+/Cl^- co-transporter^{10,11} present in pyramidal neurons. In sharp contrast to the pyramidal neurons of juvenile and adult rats, dorsal-root ganglion (DRG) cells, which are well known for their depolarizing GABA_A-receptor responses^{1,12}, showed a barely detectable level of KCC2 transcripts, whereas the expression of KCC1 was high in the DRG (Fig. 1d).

The proposal that KCC2 is necessary for fast Cl^- -dependent hyperpolarizing inhibition (see below, and Table 1) was further supported by results from northern blot analyses from hippocampi of the guinea pig, a species with precocious neonates. Here, KCC2 mRNA was present in abundant amounts already at embryonic day

42 (E42) and was not significantly upregulated during postnatal development (Fig. 1e).

To obtain anatomical information on the developmental expression patterns of KCC2, we used ³⁵S-labelled KCC2 antisense complementary RNA probes for *in situ* hybridization in sagittal sections of the developing rat brain (Fig. 2). The hippocampus and neocortex were devoid of KCC2 mRNA at E20, whereas a clear message was recorded in the thalamus (Fig. 2a, b). In the more caudal parts of the brain stem, an abundant KCC2 signal was detected already at E20 (data not shown). At P0, only a few neurons expressed KCC2 transcripts in the archi- and neocortical structures, but a stronger

Table 1 Expression of KCC2 and polarity of GABA_A-receptor-mediated responses

Preparations	KCC2 expression*	GABA _A -receptor responses
Rat hippocampus		
Pyramidal neurons, P0–P4	Low	Depolarizing ^{1–3*}
Pyramidal neurons, P > 12	High	Hyperpolarizing ^{1–3*}
Pyramidal neurons, P > 12 exposed to antisense A	Low	~0*
Pyramidal neurons, P > 12 exposed to sense A	High	Hyperpolarizing*
Adult rat DRG neurons	Low	Depolarizing ^{1,12}
Guinea-pig hippocampus		
Pyramidal neurons P0–P4	High	Hyperpolarizing*
Pyramidal neurons P > 17	High	Hyperpolarizing ^{10*}

* This study.

expression was evident around P5, and, after the first two postnatal weeks, a steady KCC2 mRNA level was attained. Thus, in gross terms, the spatiotemporal expression pattern of KCC2 mRNA appears to follow the functional maturation of the various brain areas examined¹³. This rule seemed to apply also within a given brain area as, for example, KCC2 expression was recorded in the dorsal blade of the dentate gyrus earlier than in the ventral blade (Fig. 2f). KCC2 could not be detected in *in situ* preparations from DRGs (data not shown).

Results based on the mere upregulation^{14,15} of KCC2 mRNA (Figs 1, 2) can be taken at best as suggestive, but not conclusive, evidence for a role of this transporter in the ontogeny of fast hyperpolarizing postsynaptic inhibition. Hence, we made a comparison of voltage responses evoked by muscimol, a GABA_A-receptor-specific agonist, in pyramidal neurons of both neonatal and juvenile rats and guinea pigs. These experiments demonstrated a tight link between the expression of KCC2 and the presence of GABA_A-receptor-mediated hyperpolarizing responses (Table 1). In recordings made using sharp microelectrodes with a non-Cl⁻ filling solution, iontophoretic application of muscimol (Fig. 3d) resulted in a hyperpolarization in P13–P30 pyramidal neurons in acute rat hippocampal slices, with a

mean driving force (DF_{GABA-A}) of -9.6 ± 0.9 mV ($n = 8$), where DF_{GABA-A} is defined as the difference between the reversal potential of the GABA_A-receptor-mediated response and the resting membrane potential ($DF_{GABA-A} = E_{GABA-A} - V_m$). Consistent with previous results^{2,3}, P0–P4 rat pyramidal neurons had clearly depolarizing muscimol-evoked responses ($DF_{GABA-A} = 7.0 \pm 3.3$ mV; $n = 6$). In marked contrast to this, guinea-pig neurons showed hyperpolarizing responses already at P0–P4 ($DF_{GABA-A} = -10.6 \pm 2.4$ mV; $n = 7$), which were similar to those seen in cells examined at P17–P40 ($DF_{GABA-A} = -9.0 \pm 3.3$ mV; $n = 7$).

To obtain evidence for a strictly causal link between KCC2

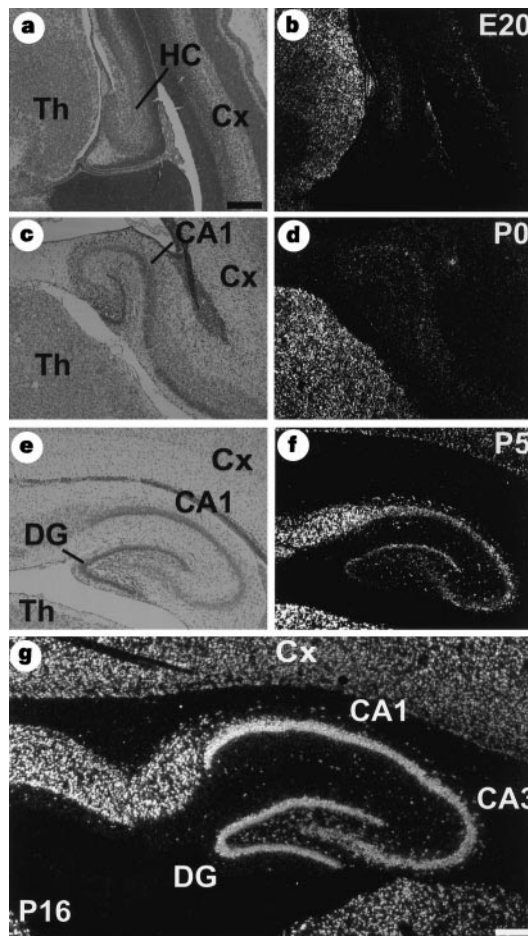


Figure 2 Expression of KCC2 mRNA in the developing rat hippocampus revealed by *in situ* hybridization. Bright-field and dark-field views of sagittal sections of late embryonic (a, b) and early postnatal (c–f) hippocampus. a, b, At E20, KCC2 mRNA labelling is seen in the thalamus (Th) but not in the hippocampal (HC) region or neocortex (Cx). c, d, At P0, weak mRNA labelling is seen in stratum pyramidale and subiculum of the hippocampus. e, f, At P5, a more intense labelling is seen in the same areas as at P0, with additional labelling of dentate gyrus (DG) and neocortex. g, A further, marked increase in labelling intensity is seen at P16 in hippocampus and neocortex. Scale bars, 200 μ m (a–f), 400 μ m (g).

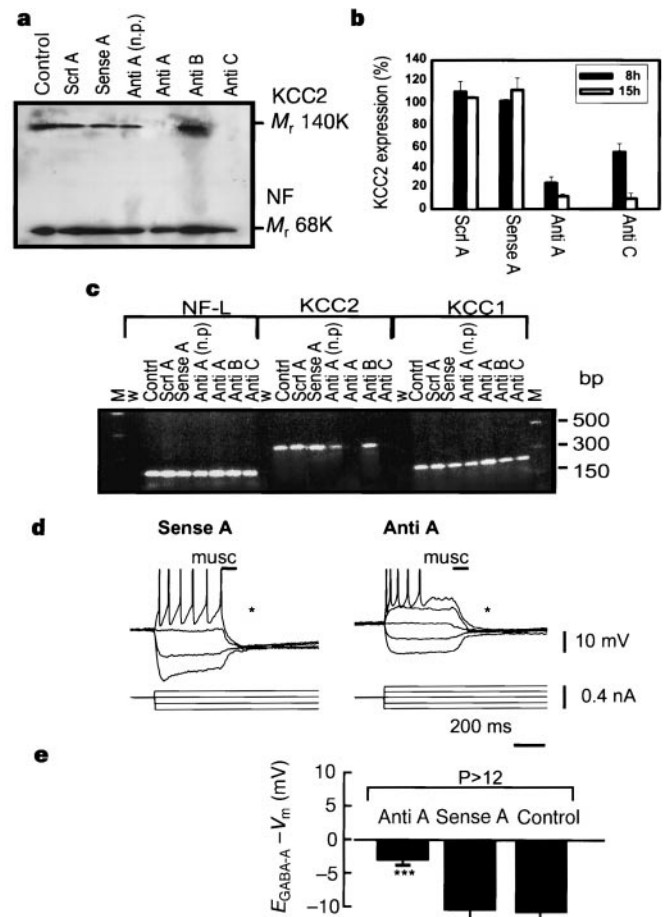


Figure 3 Specific inhibition of KCC2 protein expression by antisense-A PODN and its suppressing effect on hyperpolarizing GABA_A-receptor-mediated responses. a, Effect of a 15-h exposure to six distinct (PODNs (Scrl A; Sense A; Anti A (n.p.); Anti A; Anti B; Anti C) on KCC2 and NF-L protein expression in cultured rat hippocampal slices as seen in simultaneous western blots. Abbreviations: Control, no added oligonucleotides (ODNs); Anti A, Anti B, Anti C, three distinct phosphorothionate-protected ODNs (PODNs); Scrl A, scrambled PODN with nucleotide composition of Anti A; Sense A, PODN with sense sequence of Anti A; Anti A (n.p.), Anti A devoid of phosphorothionate protection. b, Relative protein levels ($n = 3-7$) of KCC2 after 8 h and 15 h of exposure to four PODNs (Scrl A, Sense A, Anti A, Anti C). c, Effect of the six (PODNs on KCC2, KCC1 and NF-L mRNA levels measured by RT-PCR. Note selective block of KCC2 mRNA expression by Anti A and Anti C. d, Current-clamp specimen recordings with sharp microelectrodes from short-term slice cultures exposed to Sense A and Anti A, illustrating the method of measuring E_{GABA-A} (upper horizontal bars indicate muscimol iontophoresis; asterisks indicate time of voltage data sampling). e, Pooled data from control slice cultures and those exposed to Sense A and Anti A demonstrate a significant positive shift in the driving force of GABA_A responses ($E_{GABA-A} - V_m$) caused by Anti A ($P < 0.0005$; unpaired *t*-test).

expression and hyperpolarizing GABA_A action, we devised experiments where hippocampal slices (P11–P13) were kept in culture¹⁶ for 2–5 days and exposed to 5 μ M antisense oligodeoxynucleotides (ODNs) against KCC2 mRNA, targeted against non-overlapping sequences within a region of 300 nucleotides from the starting codon. Three (A, B and C) of the four distinct 22-nucleotide antisense ODNs were phosphorothionate-protected (PODNs), and the effects of the non-protected equivalent of antisense A were also examined (Fig. 3). The (P)ODNs were 5'-labelled with fluorescein to visualize cellular uptake, which was detectable within ~1 h after exposure.

In western blot analyses from membrane preparations, the KCC2 band of relative molecular mass ~140,000 (M_r ~140K) was strongly reduced after a 15-h exposure to antisense A and C, whereas antisense B as well as the sense and scrambled counterparts of antisense A were ineffective (Fig. 3a, b). The non-protected equivalent of antisense A (also applied at 5 μ M) was much less effective than the corresponding PODN (Fig. 3a). The specificity of antisense PODN A and C action was tested in RT-PCR experiments on KCC2, KCC1 and NF-L, which showed that these were the only ODNs that had an effect on KCC2 mRNA level. None of the six ODNs tested had any influence on either KCC1 or NF-L mRNA (Fig. 3c).

Western blots from slices treated for 8 h showed that antisense A was faster than antisense C in its blocking effect on KCC2 protein expression, resulting in a reduction to $22 \pm 5\%$ ($n = 7$) in KCC2 protein expression (Fig. 3b), and it was therefore chosen for electrophysiological experiments. Measurements using ³²P-labelled antisense A indicated a final concentration of ~0.6–1.0 μ M ($n = 12$) within the slice.

All of the electrophysiological experiments on the effects of the A-type PODNs were done within an 8–15-h post-exposure time window (Fig. 3d, e). In agreement with the hypothesis that KCC2 is essential for postsynaptic hyperpolarizing inhibition, measurements of $E_{\text{GABA-A}}$ in slices treated with the antisense-A PODN consistently yielded values that were very close to the resting membrane potential (Fig. 3e), and in three of these neurons the response was even slightly depolarizing. The mean $DF_{\text{GABA-A}}$ was -2.8 ± 0.8 mV ($n = 12$). In contrast, $DF_{\text{GABA-A}}$ in slices kept in the sense-A PODN were always hyperpolarizing (-10.3 ± 1.7 mV; $n = 9$) with an amplitude similar to the control responses (-10.9 ± 0.8 mV; $n = 8$). The slightly hyperpolarizing mean $DF_{\text{GABA-A}}$ value in the antisense-A-treated slices may be attributable to a low level of KCC2 that may remain functional (Fig. 3b), or to Cl⁻ extrusion by HCO₃⁻-dependent mechanisms¹⁷, or both. There was no statistically significant difference in V_m between the two groups (pooled mean, -61.4 ± 1.1 mV; $n = 21$). Neither the antisense-A nor the sense-A PODN had any detectable effects on the passive electrical properties or spiking pattern of the pyramidal neurons.

Taken together (Table 1), these results indicate that expression of KCC2 is a necessary (although not a sufficient¹⁸) condition for fast hyperpolarizing postsynaptic inhibition. Previous studies¹⁹ have examined the role of voltage-gated Cl⁻ channels (for example, CLC-2) in the ontogeny of neuronal chloride regulation. However, although conductive systems may assist in Cl⁻ homeostasis, they cannot generate an electrochemical ionic gradient¹. As recognized previously, thermodynamics dictates that the ontogeny of fast hyperpolarizing GABAergic inhibition must be accompanied by maturation of effective net Cl⁻ extrusion^{20,21}. In light of the early expression of KCC2 in the brain stem, as seen in our *in situ* hybridization experiments, it is likely that this mechanism also sets the pace for the development of glycine-mediated fast hyperpolarizing inhibition²². This is a question for future studies to address. Moreover, changes in KCC2 expression may be important in pathological conditions such as epilepsy, and in the depolarizing shift in GABA_A responses seen in various types of neuronal trauma²³. □

Methods

RNA isolation and northern blot analysis. Total RNA was isolated from fresh rat and guinea-pig tissue of whole hippocampi by the guanidinium thiocyanate method²⁴. Northern blot analysis was done on 15 μ g of total hippocampal RNA as described⁷. A cDNA probe specific for KCC1 was prepared from a 340-base-pair (bp) *Bam*HI–*Bam*HI fragment (bp 3,387–3,726) derived from clone ERB24 (ref. 8). A cDNA probe specific for KCC2 was prepared from a 635-bp *Apa*I–*Apa*I fragment (bp 1–635) of the 5ERB14 clone⁷ for rat mRNA analysis and from the amplification product obtained from guinea-pig hippocampal cDNA (see Supplementary Information for probe sequence). Hybridization of the same membrane with an 18S rRNA specific probe was used as control.

RT-PCR from hippocampal slice cultures and single hippocampal neurons. Total RNA from rat and guinea-pig hippocampi was reverse transcribed using random hexamer primers and superscriptIII reverse transcriptase (Life Technologies Inc.). Oligodeoxynucleotide primers were synthesized over regions specific for either the KCC1 or the KCC2 cDNA. For KCC1, forward (5'-CCTGGAGTTGGGTTGTCTAAGA-3') and reverse (5'-CATCAGCCCTCACCAGTCATCTC-3') primers were used to amplify a 233-bp fragment (bp 3,278–3,510 in KCC1). For KCC2, forward (5'-CTCAAC AACCTGACGGACTG-3') and reverse (5'-GCAGAAGGACTCCATGATGCC TGCG-3') primers were used to amplify a 399-bp fragment (bp 4–402 in KCC2).

Single pyramidal neurons were isolated from P15–P20 rats²⁵ and aspirated into a presilanized glass-capillary pipette (previously kept at 200 °C overnight) containing 20 U RNase inhibitor, 0.1 μ g yeast tRNA, and 1 U RNase-free DNAase I (Boehringer) in (in mM): 120 KCl, 5 MgCl₂, 10 HEPES (pH 7.4). The cell-containing solution was collected in a 200- μ l thin-walled eppendorf tube and incubated at 37 °C for 15 min and then at 70 °C for 10 min. After reverse transcription the samples were then separated into three thin-walled tubes containing primer pairs for KCC1, KCC2 and NF-L²⁶, respectively. The samples were then subjected to 35 cycles of PCR amplification. The PCR products were separated on a 2% agarose gel and transferred to a nylon membrane (HYBOND-N+, Amersham) by semi-dry blotting. The membranes were hybridized with ³²P-labelled specific probes for each of the amplified transcripts.

Oligodeoxynucleotide-exposed hippocampal slice cultures. KCC2 antisense, scrambled or sense oligodeoxynucleotides, phosphorothionated (PODN) at all positions (except for one non-protected equivalent of antisense A) and purified by high-performance liquid chromatography²⁷, were as follows: antisense A, 5'-TCTCCTTGGGATTGCCGTA-3' (+59 relative to the ATG starting signal), antisense B, 5'-TCGCAGTCCGTCAGGTTGTT-3' (+12), antisense C, 5'-GCAGAAGGACTCCATGATGC-3' (+388); sense A, 5'-TGACGGCAATCCCAAGGAGA-3'; scrambled A, 5'-TCTTCTTGAGACTGC AGTCA-3'. The PODNs were 5'-labelled with fluorescein to visualize cellular uptake. We quantified the uptake of antisense A by measuring the uptake of the ³²P-labelled PODN by the slice cultures (34 °C).

Transverse slices of 350 μ m of the dorsal half of hippocampi from P11–P13 Wistar rats were cultured¹⁶ for 2–5 days. The slices were then exposed for 8–15 h to 5 μ M PODNs in the standard extracellular solution (in mM): 124 NaCl, 3.0 KCl, 2.0 CaCl₂, 25 NaHCO₃, 1.1 NaH₂PO₄, 2.0 MgSO₄, 10 D-glucose (34 °C). **Immunoblotting.** Rabbit polyclonal antibodies were raised against a fusion protein (termed B2C) containing most of the predicted intracellular carboxy-terminal domain of rat KCC2 (amino acids 645–1,116) (see Supplementary Information). Crude membrane fractions were prepared from whole brains and organotypic slices by differential centrifugation. Membrane proteins were separated by SDS–polyacrylamide gel electrophoresis, electrophoretically transferred to polyvinylidene difluoride (PVDF) membranes, and immunoblotted as described⁹. For antibody characterization, see Supplementary Information.

In situ hybridization. Brains of male Wistar rats were fixed intracardially with 4% paraformaldehyde in PBS and immersed in the fixative overnight. *In situ* hybridization on paraffin-embedded sections was done as described²⁸. Sagittal sections (thickness 7 μ m) were hybridized using a KCC2-specific ³⁵S-labelled antisense and sense (control) riboprobe generated from the 5ERB14-clone (nucleotides 5–834).

Electrophysiological recordings. Recordings were done from the short-term

slice cultures exposed to the PODNs (see above), as well as from acute hippocampal slices P0–P4 rats (slice thickness 600 μm), P13–P30 rats (400 μm), and P0–P4 and P17–P40 guinea pigs (400 μm) in a submerge-type chamber at 34 °C in the standard solution described above (for the neonate preparations, MgSO_4 was decreased to 1.3 mM). The intracellular sharp microelectrodes were filled with 3 M potassium methylsulphate plus 10 mM KCl, or with a mixture of potassium methylsulphate (3 M) and potassium acetate (2 M) plus 10 mM KCl, and they had a resistance of 90–150 M Ω . $E_{\text{GABA-A}}$ was measured under discontinuous current-clamp conditions. Muscimol (Tocris–Cookson) was applied iontophoretically from a micropipette filled with 100 mM muscimol in 95 mM HCl, positioned within 30–100 μm from the intracellular electrode. The amplitude and duration of the current pulses were made as low as possible (typically 10 nA, 100 ms, with 40-s inter-pulse intervals) to avoid dose-dependent shifts in $E_{\text{GABA-A}}$ (ref. 1). A backing current of –1 to –1.5 nA was used to prevent muscimol leakage from the pipette.

Received 26 October; accepted 26 November 1998.

1. Kaila, K. Ionic basis of GABA_A receptor channel function in the nervous system. *Prog. Neurobiol.* **42**, 489–537 (1994).
2. Ben-Ari, Y., Cherubini, E., Corradetti, R. & Gaiarsa, J. L. Giant synaptic potentials in immature rat CA3 hippocampal neurones. *J. Physiol. (Lond.)* **416**, 303–325 (1989).
3. Cherubini, E., Gaiarsa, J. L. & Ben-Ari, Y. GABA: an excitatory transmitter in early postnatal life. *Trends Neurosci.* **14**, 515–519 (1991).
4. Serafini, R., Valeyev, A. Y., Barker, J. L. & Poulter, M. O. Depolarizing GABA-activated Cl[–] channels in embryonic rat spinal and olfactory bulb cells. *J. Physiol. (Lond.)* **488**, 371–386 (1995).
5. Yuste, R. & Katz, L. Control of postsynaptic Ca²⁺ influx in developing neocortex by excitatory and inhibitory neurotransmitters. *Neuron* **6**, 333–344 (1991).
6. LoTurco, J., Owens, D., Heath, M., Davis, M. & Kriegstein, A. GABA and glutamate depolarize cortical progenitor cells and inhibit DNA synthesis. *Neuron* **15**, 1287–1298 (1995).
7. Payne, J. A., Stevenson, J. & Donaldson, L. Molecular characterization of a putative K–Cl cotransporter in rat brain. A neuronal-specific isoform. *J. Biol. Chem.* **271**, 16245–16252 (1996).
8. Gillen, C., Brill, S., Payne, J. A. & Forbush, B. III Molecular cloning and functional expression of the K–Cl cotransporter from rabbit, rat, and human. A new member of the cation–chloride cotransporter family. *J. Biol. Chem.* **271**, 16237–16244 (1996).
9. Payne, J. A. Functional characterization of the neuronal-specific K–Cl cotransporter KCC2: Implications for [K⁺]_o regulation. *Am. J. Physiol.* **42**, C1516–C1525 (1997).
10. Misgeld, U., Deisz, R. A., Dodt, H. U. & Lux, H. D. The role of chloride transport in postsynaptic inhibition of hippocampal neurons. *Science* **232**, 1413–1415 (1986).
11. Thompson, S. M. & Gähwiler, B. H. Activity-dependent disinhibition. II. Effect of extracellular potassium, furosemide, and membrane potential on E_{Cl} in hippocampal CA3 neurons. *J. Neurophysiol.* **61**, 512–523 (1989).
12. Deschenes, M., Feltz, P. & Lamour, Y. A model for an estimate in vivo of the ionic basis of presynaptic inhibition: an intracellular analysis of the GABA-induced depolarization in rat dorsal root ganglia. *Brain Res.* **118**, 486–493 (1976).
13. Bayer, S. A. & Altman, J. Directions in neurogenetic gradients and patterns of anatomical connections in the telencephalon. *Prog. Neurobiol.* **29**, 57–106 (1987).
14. Rivera, C., Wegelius, K., Reeb, M., Saarma, M. & Kaila, K. Developmental regulation of K–Cl cotransporter (KCC2 and KCC1) mRNA in early postnatal rat hippocampus. *Soc. Neurosci. Abstr.* **23**, 26.7 (1997).
15. Clayton, G. H., Owens, G. C., Wolff, J. S. & Smith, R. L. Ontogeny of cation–Cl cotransporter expression in rat neocortex. *Brain Res. Dev. Brain Res.* **109**, 281–292 (1998).
16. Stoppini, L., Buchs, P. A. & Müller, D. A simple method for organotypic cultures of nervous tissue. *J. Neurosci. Methods* **37**, 173–182 (1991).
17. Bevensee, M. O. & Boron, W. F. In *pH and Brain Function* (eds Kaila, K. & Ransom, B.) 211–231 (Wiley, New York, 1998).
18. Kaila, K., Voipio, J., Paalasmaa, P., Pasternack, M. & Deisz, R. A. The role of bicarbonate in GABA_A receptor-mediated IPSPs of rat neocortical neurons. *J. Physiol. (Lond.)* **464**, 273–289 (1993).
19. Clayton, G. H., Staley, K. J., Wilcox, C. L., Owens, G. C. & Smith, R. L. Developmental expression of CLC-2 in the rat nervous system. *Brain Res. Dev. Brain Res.* **108**, 307–318 (1998).
20. Zhang, L., Spiegelman, I. & Carlen, P. Development of GABA-mediated chloride-dependent inhibition in CA1 pyramidal neurones of immature rat hippocampal slices. *J. Physiol. (Lond.)* **444**, 25–49 (1991).
21. Sutor, B. & Luhmann, H. Development of excitatory and inhibitory postsynaptic potentials in the rat neocortex. *Perspectives Dev. Neurobiol.* **2**, 409–419 (1994).
22. Wu, W. L., Ziskind-Conhaim, L. & Sweet, M. A. Early development of glycine and GABA mediated synapses in rat spinal cord. *J. Neurosci.* **12**, 3935–3945 (1992).
23. van den Pol, A. N., Obrietan, K. & Chen, G. Excitatory action of GABA after neuronal trauma. *J. Neurosci.* **16**, 4283–4292 (1996).
24. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
25. Pasternack, M., Smirnov, S. & Kaila, K. Proton modulation of functionally distinct GABA_A receptors in acutely isolated pyramidal neurons of rat hippocampus. *Neuropharmacology* **35**, 1279–1288 (1996).
26. Moshnyakov, M., Arumäe, U. & Saarma, M. mRNAs for one two or three members of trk receptor family are expressed in single rat trigeminal ganglion neurons. *Mol. Brain Res.* **43**, 141–148 (1996).
27. Bengström, M. & Paulin, L. Synthesis and purification of thio-oligonucleotides. *Nucleic Acids Symp. Ser.* **24**, 288 (1991).
28. Pirvola, U. et al. Brain-derived neurotrophic factor and neurotrophin 3 mRNAs in the peripheral target fields of developing inner ear ganglia. *Proc. Natl Acad. Sci. USA* **89**, 9915–9919 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank F. Edwards, T. Freund and H. Rauvala for comments on this paper. This work was supported by the Academy of Finland (K.K. and M.S.), the Sigrid Juselius Foundation (K.K. and M.S.), the Swedish Council for Natural Sciences (C.R.) and Hibbard E. Williams Research Fund, National Institute of Neurological Disorders and Stroke, Epilepsy Foundation of America (J.A.P.).

Correspondence and requests for materials should be addressed to K.K. (e-mail: kai.kaila@helsinki.fi).

Polyunsaturated fatty acids activate the *Drosophila* light-sensitive channels TRP and TRPL

Sylwester Chyb*, Padinjat Raghu & Roger C. Hardie

Cambridge University Department of Anatomy, Downing St, Cambridge CB2 3DY, UK

* Present address: Department of Biochemistry, Cambridge University, Tennis Court Road, Cambridge CB2 1GA, UK.

Phototransduction in invertebrate microvillar photoreceptors is thought to be mediated by the activation of phospholipase C (PLC), but how this leads to gating of the light-sensitive channels is unknown^{1,2}. Most attention has focused on inositol-1,4,5-trisphosphate, a second messenger produced by PLC from phosphatidylinositol-4,5-bisphosphate; however, PLC also generates diacylglycerol, a potential precursor for several polyunsaturated fatty acids, such as arachidonic acid and linolenic acid. Here we show that both of these fatty acids reversibly activate native light-sensitive channels (transient receptor potential (TRP) and TRP-like (TRPL)) in *Drosophila* photoreceptors as well as recombinant TRPL channels expressed in *Drosophila* S2 cells. Recombinant channels are activated rapidly in both whole-cell recordings and inside-out patches, with a half-maximal effector concentration for linolenic acid of ~10 μM . Four different lipoxygenase inhibitors, which might be expected to lead to build-up of endogenous fatty

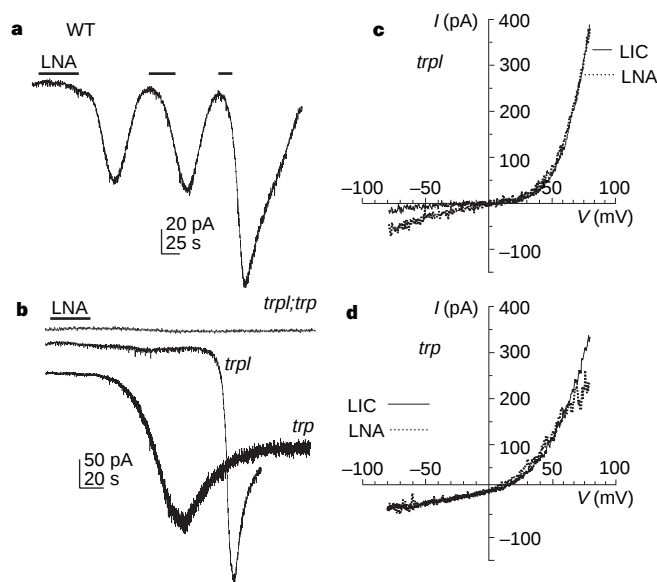


Figure 1 Activation of light-sensitive channels by linolenic acid (LNA). **a**, Whole-cell recording from a wild-type (WT) photoreceptor at –70 mV in the dark. 20 μM LNA induced a noisy inward current which could be activated repeatedly. **b**, LNA was similarly effective in both *trpl* and *trp*^{CM} mutants, but no current was activated in the double mutant *trpl*; *trp*^{CM}. Greater channel noise in *trp*^{CM} reflects the ten times greater single-channel conductance of the TRPL channels⁸. **c, d**, Current–voltage relationship for light-induced conductance (LIC) and LNA-activated conductance (LNA) in *trpl* (**c**) and *trp* (**d**) mutants, determined by applying a voltage ramp from –80 to +80 mV and subtracting a template recorded in the dark; the slopes were scaled by eye to overlap. The dual inward/outward rectification of TRP channels (in *trpl* mutants) and the simple outward rectification of TRPL channels (in *trp* mutants) are characteristic features of these conductances⁸.

acids, also activate native TRP and TRPL channels in intact photoreceptors. As arachidonic acid may not be found in *Drosophila*, we suggest that another polyunsaturated fatty acid, such as linolenic acid, may be a messenger of excitation in *Drosophila* photoreceptors.

In *Drosophila* photoreceptors, which have been widely used as a model system for analysing both invertebrate phototransduction and phosphoinositide signalling^{1–3}, activation of PLC is coupled to the opening of two classes of light-sensitive channel, which are, at least partly, encoded by *trp* and *trpl*^{4–8}. Vertebrate TRP homologues have also been discovered and may be involved in a variety of receptor-mediated Ca²⁺-influx phenomena^{9–11}. Despite intensive investigation, the mechanism of activation of both the *Drosophila* channels and their vertebrate homologues has remained unknown. PLC generates not only inositol-1,4,5-trisphosphate (InsP₃) but also diacylglycerol (DAG), but this possibility has been largely ignored as a possible route for excitation in *Drosophila* because an immediate target of DAG, protein kinase C (PKC), appears to be involved only in response inactivation and light adaptation^{12,13}. However, DAG is also a potential precursor for polyunsaturated fatty acids (PUFAs) such as arachidonic acid; PUFAs can be released from DAG by the action of DAG lipase. Arachidonic acid may be absent in *Drosophila*¹⁴, but related PUFAs, including linolenic and linoleic acids, are present¹⁵ and should be released by the same enzymes. We found that arachidonic, linolenic and linoleic acids have similar effects (see below), but we suggest that linolenic or linoleic acids may be physiologically relevant in *Drosophila*.

Whole-cell recordings from dissociated wild-type *Drosophila* photoreceptors can usually be maintained with a stable baseline indefinitely (>30 mins), but when linolenic or arachidonic acid (20 μ M) was applied from a nearby puffer pipette the light-sensitive channels were always activated ($n = 11$ cells for linolenic acid; $n = 8$ cells for arachidonic acid) in a reversible manner after a

Table 1 TRPL-channel properties in inside-out patches from S2 cells

	pS	Mean open time (ms)	τ_1 (ms)	τ_2 (ms)	w_1/w_2
Spontaneous	66 $\pm$ 4	0.98 $\pm$ 0.22	0.29 $\pm$ 0.07	1.3 $\pm$ 0.4	4.5 $\pm$ 1.7
LNA	69 $\pm$ 6	1.02 $\pm$ 6	0.35 $\pm$ 0.11	2.0 $\pm$ 1.2	3.6 $\pm$ 0.7

Single-channel conductances (pS), mean open times, time constants of two exponential fits to open time distributions (τ_1 , τ_2) and the relative weighting of these time constants (w_1 , w_2), before (spontaneous) and after activation by 10 μ M linolenic acid (LNA), were statistically indistinguishable (Student's *t*-test). Spontaneously active channels have been shown previously to be indistinguishable from native TRPL channels²⁰. Data are based on ~1,000 idealized channel openings per patch, presented as mean $\pm$ s.d. across patches ($n = 11$), recorded at +60 mV, sampled at 10 kHz, and filtered at 2 kHz.

delay of several seconds (Fig. 1a; only the response to linolenic acid is shown in Fig. 1). The current was composed initially of noisy fluctuations, which subsequently developed into an overshooting inward current with high-frequency channel noise, during which time the response to light was markedly, but reversibly, attenuated. In similar experiments using *trp* ($n = 6$) and *trpl* ($n = 10$) mutants to isolate TRPL- and TRP-dependent channels, respectively, both classes of channel were always activated by similar doses of arachidonic acid or linolenic acid (Fig. 1b). Identification of the conductances was confirmed by their current–voltage relationships, which were similar to those for light-induced conductances in the same cells (Fig. 1c, d), and by the fact that no channels could be activated in a *trpl*; *trp* double mutant ($n = 6$).

PUFAs can activate PKC, PLC and G proteins¹⁶. To exclude the possibility that arachidonic or linolenic acids were activating TRP and TRPL through any of these routes, we made recordings from photoreceptors in *norpa*^{P24}, *inaC*²⁰⁹ and *Gaq*¹ mutants, which are null or near-null mutants of photoreceptor PLC, PKC and G_q, respectively^{12,17,18}. Linolenic acid (20 μ M) activated the light-sensitive channels in all photoreceptors tested in all three mutants ($n = 3–4$ cells in each) with a similar potency, indicating that it may act downstream of G proteins and PLC.

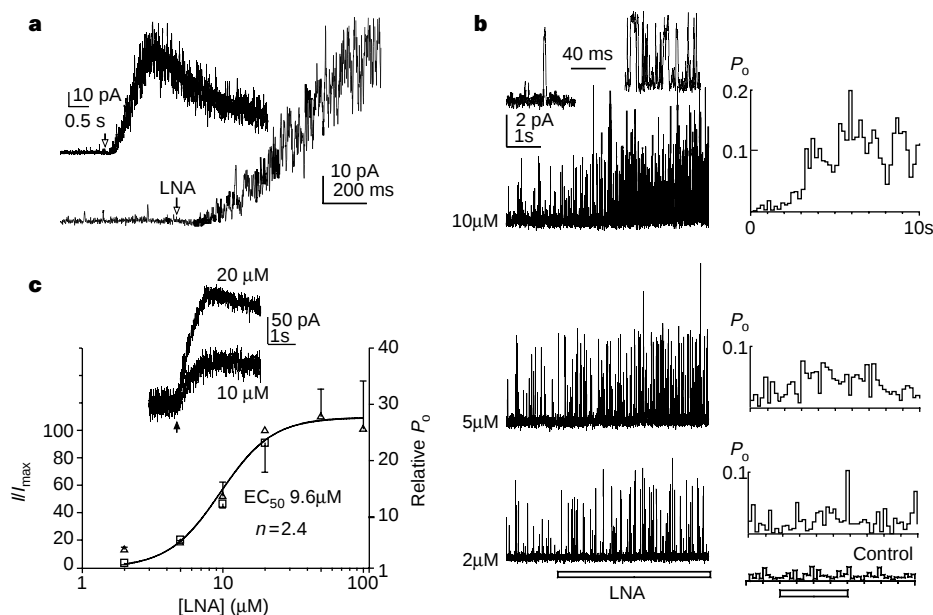


Figure 2 Activation of recombinant TRPL channels by linolenic acid (LNA). **a**, Whole-cell recording from an S2 cell expressing *trpl* cDNA clamped at +50 mV; 20 μ M LNA (applied for 1 s, arrow; not corrected for perfusion delay) induced rapid channel activation. The same response is shown on two different scales. **b**, Left, recordings from excised inside-out patches containing 2–3 channels held at +60 mV and exposed to different concentrations of LNA (10, 5 or 2 μ M). Representative openings before and after application of 10 μ M LNA are shown at the top on a faster timescale. Right, time course of channel open probability (P_o) calculated in 200-ms bins from idealized channel openings from these traces, as

well as for a control obtained using 25 μ M stearic acid. **c**, Dose-response function for LNA, determined in whole-cell recordings (triangles) and excised inside-out patches (squares). Whole-cell data (I/I_{max}) were obtained by applying two doses (20 μ M reference plus one another) alternatively from a double puffer pipette (representative traces are shown in the inset); single-channel data are expressed as an increase in open probability (relative P_o) relative to the spontaneous open probability before LNA application. Data are shown as mean $\pm$ s.e.m.; $n = 4–7$ cells or patches for each point. Data are fitted by a Hill plot, with an effective EC₅₀ of 9.6 μ M and a Hill coefficient of 2.4.

Table 2 Agonist profile of TRPL channels in whole-cell recordings from S2 cells

	Arachidonic*	Linoleic	Oleic	Fatty acid			
	20:4 <i>cis</i>	18:2 <i>cis</i>	18:1 <i>cis</i>	ETI*	Linolenelaidic*	Linolelaidic	Elaidic
					18:3 <i>trans</i>	18:2 <i>trans</i>	18:1 <i>trans</i>
Potency	0.9 ± 0.2	1.0 ± 0.2	0.2 ± 0.1	0.9 ± .07	0.5 ± 0.1	0.4 ± 0.03	0.03 ± 0.07

Potency is defined as the response to a 1-s 20 μ M application of fatty acid relative to the response to 20 μ M linolenic acid (LNA) elicited with the other barrel of a double puffer pipette (mean $\pm$ s.e.m., $n = 4$ –7 cells).

* These substances were also tested in inside-out patches ($n = 3$), resulting in a significant increase (≥ 5 -fold) in the TRPL-channel open probability in each case. Control substances giving no response included the unsaturated fatty acids stearic and palmitic acid (up to 100 μ M tested; $n = 4$ cells, 4 inside-out patches) and InsP₃ (100 μ M tested; $n = 7$ patches).

The relatively slow onset of activation of TRP and TRPL might reflect an indirect action of PUFAs; however, other possible explanations are that access to the microvillar membrane might be restricted, or that PUFA concentrations could be controlled by endogenous inactivation mechanisms or buffered by a fatty-acid-binding glycoprotein found in the intra-ommatidial matrix¹⁹. We therefore also made recordings from TRPL channels heterologously expressed in *Drosophila* S2 cells; channel identity was confirmed by the characteristic biophysical properties of the expressed channels (Table 1) which seem to be indistinguishable from the properties of native TRPL channels²⁰, and by the absence of any such channels in untransfected cells. Bath application of linolenic acid or arachidonic acid in both whole-cell recordings ($n > 100$) and excised inside-out patches ($n = 70$) always activated recombinant TRPL channels, with a steep dose–response function (threshold for linolenic acid ~ 2 μ M; effector concentration for half-maximal response (EC₅₀) ~ 10 μ M; Fig. 2a–c). After correction for perfusion times, latencies in whole-cell recordings were negligible (0–50 ms). Although latencies in patches could be only roughly estimated because of spontaneous activity and the stochastic nature of channel openings, activation appeared somewhat slower than in whole-cell recordings (~ 0.1 –1 s delays); however, reversible activation by linolenic acid applied to the cytoplasmic surface of inside-out patches was observed up to at least 30 min after patch excision, strongly suggesting a direct effect of linolenic acid.

Fatty-acid effects on ion channels have sometimes been attributed to nonspecific effects on membranes¹⁶. The effects of a nonspecific detergent should be mimicked by saturated fatty acids (for example, palmitic and stearic acid), but neither influenced channel activity in S2 cells (whole-cell or inside-out patches) or photoreceptors (at concentrations of fatty acids of up to 100 μ M; Fig. 2b). *Cis*-unsaturated fatty acids, such as linolenic acid and arachidonic acid, would be expected to increase membrane fluidity. This effect should not be mimicked by *trans*-isomers^{16,21}, which can, however, act as agonists²². As *trans*-isomers, including linolenelaidic and linolelaidic acid (20 μ M), also activated TRPL channels in whole-cell and inside-out-patch recordings from S2 cells (Table 2) as well as TRP channels in *trpl* mutant photoreceptors ($n = 3$), a change in membrane fluidity was probably not required for channel activation. TRPL channels were also activated by a mono-unsaturated fatty acid (oleic acid, 20 μ M; $n = 4$ cells) and the arachidonic acid analogue 5,8,11-eicosatriynoic acid (ETI, 20 μ M; $n = 5$ cells, 3 patches), neither of which are substrates for the major PUFA-metabolizing enzymes (cyclo-oxygenase and lipoxygenase). These results further support the idea that PUFAs act directly rather than, for example, through one of the many possible PUFA metabolites. Overall the agonist profile for the TRPL channels studied here (Table 2) is typical of ion channels that are directly modulated by fatty acids^{16,22} and, taken together, these results indicate that activation of TRP and TRPL channels may occur by direct ligand interaction with the channel protein.

If a PUFA is involved in excitation, one would expect its concentration to be controlled by endogenous inactivation mechanisms. We therefore tested inhibitors of enzymes that metabolize PUFAs in the expectation that these might lead to uncontrolled build-up of endogenous PUFA in the photoreceptors. Although the

cyclo-oxygenase inhibitor indomethacin (100 μ M) had no effect, brief application of the lipoxygenase inhibitor nordihydroguaiaretic acid (NDGA) (50 μ M) rapidly induced discrete events that were similar in waveform to quantum bumps (Fig. 3d). Such events, normally induced by the absorption of single photons, are thought to be shaped by rapid, sequential, positive and negative feedbacks by Ca²⁺ influx once the amount of endogenous transmitter reaches a critical threshold; the ability of NDGA to mimic such responses is indicative of activation by a physiological route. With further application of NDGA the bump-like events increased in frequency, leading to an overshooting, noisy inward current (Fig. 3a, b). Responses to light flashes during the onset of the inward current appeared to have a specific defect in deactivation, as would be predicted if lipoxygenase action were involved in response inactivation (Fig. 3c). NDGA activated channels in all photoreceptors tested ($n > 30$) at concentrations as low as 10 μ M. Similar results were obtained with another selective lipoxygenase inhibitor, cinnamyl-3,4-dihydroxy- α -cyanocinnamate (CDC, ≥ 10 μ M), and with ETI (50 μ M) and 5,8,11,14-eicosatetraynoic acid (ETYA, ≥ 10 μ M) which are competitive analogues of arachidonic acid and inhibit both lipoxygenase and other enzymes that use arachidonic acid as substrates. At low doses the effects of the inhibitors were reversible.

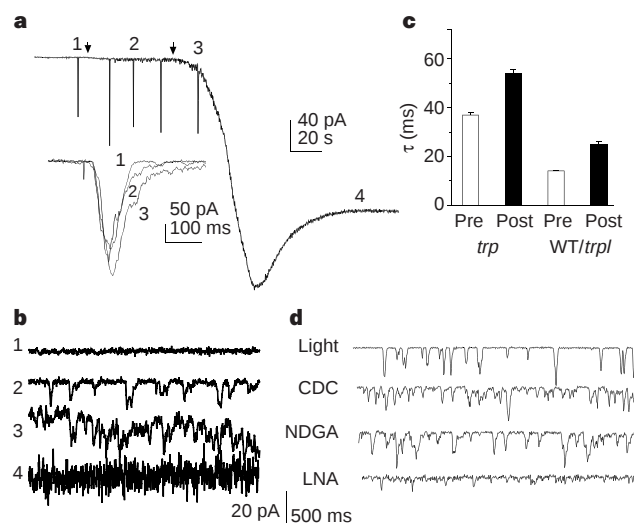


Figure 3 Activation of photoreceptors by lipoxygenase inhibitors. **a**, A 5-s application (first arrow) of NDGA (50 μ M) induced discrete bump-like events in the dark in a wild-type photoreceptor; following a further 5-s dose (second arrow), these bump-like events accelerated, generating an overshooting inward current. Inset, responses to 10-ms flashes (indicated by stimulus artefact) recorded during the onset of the current showed a specific defect in deactivation. **b**, Selected segments of the overshooting inward current from **a** shown on an expanded scale. **c**, Time constants of decay (τ) of flash responses from single exponential fits before (pre) and after (post) application of NDGA in *trp* mutants (mean $\pm$ s.e.m.; $n = 11$) and wild-type or *trpl* flies (mean $\pm$ s.e.m.; $n = 7$). **d**, Bump-like events induced by CDC and NDGA (brief, 50 μ M application) were similar in size and duration to light-induced quantum bumps; LNA induced smaller, noisy fluctuations.

At similar doses, all four inhibitors activated TRP-dependent channels isolated in *trpl* mutant photoreceptors; NDGA, ETI and ETYA also activated TRPL channels in *trp* mutants. However, CDC had the additional effect of blocking TRPL channels (see below), and, at best, a low level of channel activation by CDC could be detected in *trp* mutants, presumably because of the competing effects of channel activation and inhibition. NDGA had no effect on TRPL channels expressed in S2 cells, indicating that it did not act at the level of the channels themselves and that lipoxygenase may be particularly important for controlling PUFA levels in photoreceptors while playing little, if any, role in the cell lines. However, as mentioned, ETI appeared to act as a direct agonist and also activated TRPL channels in S2 cells.

As there are no reported antagonists of TRPL channels, it was interesting that CDC, one of the lipoxygenase inhibitors that activated channels in wild-type flies and *trpl* mutants, specifically blocked the response to light in *trp* mutants but not in wild-type flies or *trpl* mutants. CDC also rapidly and reversibly blocked TRPL channels in S2 cells with a half-maximal inhibitory concentration of $\sim 1 \mu\text{M}$ (Fig. 4). CDC and ETI are, to our knowledge, the first reported pharmacological agents that affect TRPL channels and it may be significant that their normal targets (such as lipoxygenase) are enzymes that bind (and metabolize) PUFAs.

In conclusion, we have shown that *Drosophila* light-sensitive channels may be activated directly by PUFAs and activated indirectly by lipoxygenase inhibitors. Together with the new pharmacology of the TRPL channels, these results lead us to suggest that a PUFA such as linolenic acid may be an excitatory messenger of phototransduction in *Drosophila*. Although this possibility need not exclude an effect of InsP_3 in phototransduction (for example, linolenic acid might be produced by a Ca^{2+} -dependent phospholipase A_2 rather than by DAG lipase), there has been no demonstration of an excitatory effect of InsP_3 in *Drosophila* photoreceptors and a mutation in the InsP_3 receptor has been found to have no effect on phototransduction²³. Although InsP_3 has been

reported to increase the activity of heterologously expressed TRPL channels²⁴, we and others^{25,26} have been unable to confirm this result. Interestingly, receptor-mediated Ca^{2+} influx mediated by arachidonic acid has been described^{27,28}. Although the channels responsible for this Ca^{2+} influx have not been molecularly characterized, they may include one or more of the growing family of vertebrate TRP homologues. □

Methods

Cell lines. *Drosophila* S2 cells transfected with *trpl* and *DM1* (*Drosophila* muscarinic receptor gene) complementary DNA and clonally selected were maintained in Shields & Sang M3 growth medium (Sigma) containing 12.5% fetal calf serum as described²⁰. Expression of DM1 and TRPL was controlled by the metallothionein promoter, which was activated by, 24–48 h before recording, adding CuSO_4 (0.6 mM) to the growth medium of cells that had just reached confluence.

Flies. The 'wild-type' strain was *Drosophila melanogaster* Oregon w. To isolate TRPL channels we used *w;trp³⁰¹* or *w;trp^{CM}* strains, which are null or near-null mutants that lack nearly all TRP protein according to functional assays^{4,8}. To isolate TRP-dependent channels we used the *trpl³⁰²,cn,bw* strain, a null mutant provided by C. Zuker⁷. To reduce or eliminate photoreceptor PLC, PKC and G_q , we used *w,norpa^{P24}* and *w;inaC^{P209}* strains (from W. L. Pak) and the *w;Gαq¹* strain (from C. Zuker); these are the most severe reported mutants in each case^{12,17,18}. Flies were raised at 25 °C and used within 2 h of eclosion.

Electrophysiology. Whole-cell recordings were performed on dissociated *D. melanogaster* ommatidia and on S2 cells as described^{20,29,30}. Because of the strong outward rectification of TRPL channels, we used positive holding potentials to enhance signals in S2 cells. Inside-out patches from S2 cells were made using standard techniques and Sylgard electrodes. Both whole-cell and single-channel data were acquired and analysed using pClamp 6.0 (Axon Instruments).

Solutions and chemicals. Physiological Ringer's solution for photoreceptor recordings contained (in mM): 120 NaCl, 5 KCl, 4 MgCl_2 , 1.5 CaCl_2 , 10 *N*-Tris-(hydroxymethyl)-methyl-2-amino-ethanesulphonic acid (TES), 25 proline and 5 alanine. For whole-cell recordings from S2 cells, the bath contained (in mM): 130 sodium gluconate, 10 TES, 4 MgCl_2 , 1.5 CaCl_2 and 30 mannitol. The standard pipette solution for photoreceptors contained (in mM): 140 potassium gluconate, 10 TES, 2 MgCl_2 , 4 Mg-ATP, 0.4 Na-GTP and 1 NAD. To analyse current-voltage relationships, we used a pipette containing (in mM): 120 CsCl, 15 TEA and 10 TES (plus nucleotide additives). For S2 cells the nucleotide additives were omitted. To study inside-out patches (S2 cells only) we used a bath solution containing (in mM): 130 sodium gluconate, 2 MgCl_2 , 10 TES, with Ca^{2+} buffered to below 100 nM with 100–200 μM EGTA; the pipette solution was the same except for the addition of 1.5 mM CaCl_2 . Drugs were applied through one or two puffer pipettes (outside diameter $\sim 2 \mu\text{m}$) positioned $\sim 50 \mu\text{m}$ from the recorded cells or patches and controlled by a WPI picopump. Perfusion times were estimated using ionic substitution and latencies were estimated from the intersection of linear fits to traces immediately before and after application. Arachidonic acid (Sigma) was used as a sodium salt from a 10 mM aqueous stock. Other fatty acids were obtained from Sigma or Biomol and maintained as 10 mM or 20 mM stock solutions in ethanol or methanol. Stock solutions of other agents were as follows: NDGA and indomethacin, 50 mM in ethanol; CDC, ETI and ETYA, 10 mM. Working solutions were made fresh daily from stocks kept at -20°C . The maximum concentration of each solvent (dimethyl sulfoxide 0.2%; ethanol and methanol 0.5%) used in any experiment was also tested alone in controls and had no effects in either S2 cells or photoreceptors.

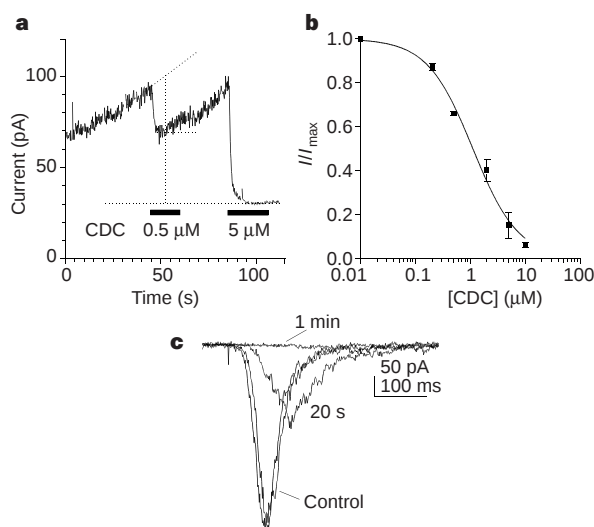


Figure 4 CDC is a new inhibitor of TRPL channels. **a**, Spontaneous TRPL activity recorded from an S2 cell held at +50 mV was blocked partially by 0.5 μM CDC and completely by 5 μM CDC. Because spontaneous TRPL activity increases continuously during whole-cell recording²⁰, the degree of block was estimated using a linear extrapolation of the current immediately before application of CDC. **b**, Dose-response function of inhibition by CDC determined from measurements as described in **a** (mean $\pm$ s.e.m.; $n = 4$ –7 cells for each point). Data were fitted by a single-site-binding function with an IC_{50} of 1 μM . **c**, Responses to 10-ms light flashes recorded from a *trp* mutant photoreceptor before (control) and 20 s and 1 min after onset of puffer application of 50 μM CDC.

Received 2 September; accepted 17 November 1998.

- Hardie, R. C. & Minke, B. Phosphoinositide-mediated phototransduction in *Drosophila* photoreceptors: the role of Ca^{2+} and *trp*. *Cell Calcium* **18**, 256–274 (1995).
- Ranganathan, R., Malicki, D. M. & Zuker, C. S. Signal transduction in *Drosophila* photoreceptors. *Annu. Rev. Neurosci.* **18**, 283–317 (1995).
- Berridge, M. J. Inositol trisphosphate and calcium signalling. *Nature* **361**, 315–325 (1993).
- Montell, C. & Rubin, G. M. Molecular characterization of *Drosophila trp* locus, a putative integral membrane protein required for phototransduction. *Neuron* **2**, 1313–1323 (1989).
- Phillips, A. M., Bull, A. & Kelly, L. E. Identification of a *Drosophila* gene encoding a calmodulin-binding protein with homology to the *trp* phototransduction gene. *Neuron* **8**, 631–642 (1992).
- Hardie, R. C. & Minke, B. The *trp* gene is essential for a light-activated Ca^{2+} channel in *Drosophila* photoreceptors. *Neuron* **8**, 643–651 (1992).

7. Niemeyer, B. A., Suzuki, E., Scott, K., Jalink, K. & Zuker, C. S. The *Drosophila* light-activated conductance is composed of the two channels TRP and TRPL. *Cell* **85**, 651–659 (1996).
8. Reuss, H., Mojet, M. H., Chyb, S. & Hardie, R. C. *In vivo* analysis of the *Drosophila* light-sensitive channels, TRP and TRPL. *Neuron* **19**, 1249–1259 (1997).
9. Petersen, C. C. H., Berridge, M. J., Borge, M. F. & Bennett, D. L. Putative capacitative calcium entry channels: expression of *Drosophila* trp and evidence for the existence of vertebrate homologues. *Biochem. J.* **311**, 41–44 (1995).
10. Wes, P. D. et al. TRPC1, a human homolog of a *Drosophila* store-operated channel. *Proc. Natl Acad. Sci. USA* **92**, 9652–9656 (1995).
11. Birnbaumer, L. et al. On the molecular basis and regulation of cellular capacitative calcium entry: roles for Trp proteins. *Proc. Natl Acad. Sci. USA* **93**, 15195–15202 (1996).
12. Smith, D. P. et al. Photoreceptor deactivation and retinal degeneration mediated by a photoreceptor-specific protein-kinase-C. *Science* **254**, 1478–1484 (1991).
13. Hardie, R. C. et al. Protein kinase C is required for light adaptation in *Drosophila* photoreceptors. *Nature* **363**, 634–637 (1993).
14. Yoshioka, T. et al. Evidence that arachidonic acid is deficient in phosphatidylinositol of *Drosophila* heads. *J. Biochem.* **98**, 657–662 (1986).
15. Stark, W. S., Lin, T. N., Brackmann, D., Christianson, J. S. & Sun, G. Y. Fatty-acids in the lipids of *Drosophila* heads—effects of visual mutants, carotenoid deprivation and dietary fatty-acids. *Lipids* **28**, 345–350 (1993).
16. Meves, H. Modulation of ion channels by arachidonic acid. *Prog. Neurobiol.* **43**, 175–186 (1994).
17. Pearn, M. T., Randall, L. L., Shortridge, R. D., Burg, M. G. & Pak, W. L. Molecular, biochemical, and electrophysiological characterization of *Drosophila* norpA mutants. *J. Biol. Chem.* **271**, 4937–4945 (1996).
18. Scott, K., Becker, A., Sun, Y., Hardy, R. & Zuker, C. G(q alpha) protein function *in vivo*: genetic dissection of its role in photoreceptor cell physiology. *Neuron* **15**, 919–927 (1995).
19. Shim, K. et al. Control of *Drosophila* retinoid and fatty acid binding glycoprotein expression by retinoids and retinoic acid: northern, western and immunocytochemical analyses. *Exp. Eye Res.* **65**, 717–727 (1997).
20. Hardie, R. C., Reuss, H., Lansdell, S. J. & Millar, N. S. Functional equivalence of native light-sensitive channels in the *Drosophila* trp³⁰¹ mutant and TRPL cation channels expressed in a stably transfected *Drosophila* cell line. *Cell Calcium* **21**, 431–440 (1997).
21. Karnovsky, M. J., Kleinfeld, A. M., Hoover, R. L. & Klausner, R. D. The concept of lipid domains in membranes. *J. Cell Biol.* **94**, 1–6 (1982).
22. Anderson, M. P. & Welsh, M. J. Fatty acids inhibit apical membrane chloride channels in airway epithelia. *Proc. Natl Acad. Sci. USA* **87**, 7334–7338 (1990).
23. Acharya, J. K., Jalink, K., Hardy, R. W., Hartenstein, V. & Zuker, C. S. InsP₃ receptor is essential for growth and differentiation but not for vision in *Drosophila*. *Neuron* **18**, 881–887 (1997).
24. Dong, Y., Kunze, D. L., Vaca, L. & Schilling, W. P. Ins(1,4,5)P₃ activates *Drosophila* cation channel Trp1 in recombinant baculovirus-infected Sf9 insect cells. *Am. J. Physiol. Cell Physiol.* **269**, C1332–C1339 (1995).
25. Zimmer, S., Trost, C., Cavalié, A., Philipp, S. & Flockerkzi, V. The trpl protein is a Ca²⁺-calmodulin activated non-selective cation (CAN) channel. *Naunyn-Schmiedeberg Arch. Pharmacol.* **355**, 238 (1997).
26. Obukhov, A. G. et al. Direct activation of trpl cation channels by G alpha11 subunits. *EMBO J.* **15**, 5833–5838 (1996).
27. Van der Zee, L., Nelemans, A. & Den Hertog, A. Arachidonic acid is functioning as a second messenger in activating the Ca²⁺ entry process on H1-histaminergic stimulation in DDT1 MF-2 cells. *Biochem. J.* **305**, 859–864 (1995).
28. Shuttleworth, T. J. Arachidonic acid activates the noncapacitative entry of Ca²⁺ during [Ca²⁺]_i oscillations. *J. Biol. Chem.* **271**, 21720–21725 (1996).
29. Hardie, R. C. Whole-cell recordings of the light-induced current in *Drosophila* photoreceptors: evidence for feedback by calcium permeating the light sensitive channels. *Proc. R. Soc. Lond. B* **245**, 203–210 (1991).
30. Hardie, R. C. INDO-1 measurements of absolute resting and light-induced Ca²⁺ concentration in *Drosophila* photoreceptors. *J. Neurosci.* **16**, 2924–2933 (1996).

Acknowledgements. We thank C. Taylor and Z. Selinger for discussions, and L. Kelly and W. A. Harris for discussions and comments on an earlier version of this paper. The work was supported by the Wellcome Trust and the BBSRC.

Correspondence and requests for materials should be addressed to R.C.H. (e-mail: rch14@hermes.cam.ac.uk).

Direct activation of human TRPC6 and TRPC3 channels by diacylglycerol

Thomas Hofmann, Alexander G. Obukhov, Michael Schaefer, Christian Harteneck, Thomas Gudermann & Günter Schultz

Institut für Pharmakologie, Universitätsklinikum Benjamin Franklin, Freie Universität Berlin, Thielallee 69-73, 14195 Berlin, Germany

Eukaryotic cells respond to many hormones and neurotransmitters with increased activity of the enzyme phospholipase C and a subsequent rise in the concentration of intracellular free calcium ([Ca²⁺]_i). The increase in [Ca²⁺]_i occurs as a result of the release of Ca²⁺ from intracellular stores and an influx of Ca²⁺ through the plasma membrane^{2–4}; this influx of Ca²⁺ may⁵ or may not⁶ be store-dependent. *Drosophila* transient receptor potential (TRP) pro-

teins and some mammalian homologues (TRPC proteins) are thought to mediate capacitative Ca²⁺ entry^{7–9}. Here we describe the molecular mechanism of store-depletion-independent activation of a subfamily of mammalian TRPC channels. We find that hTRPC6 is a non-selective cation channel that is activated by diacylglycerol in a membrane-delimited fashion, independently of protein kinases C activated by diacylglycerol. Although hTRPC3, the closest structural relative of hTRPC6, is activated in the same way, TRPCs 1, 4 and 5 and the vanilloid receptor subtype 1 are unresponsive to the lipid mediator. Thus, hTRPC3 and hTRPC6 represent the first members of a new functional family of second-messenger-operated cation channels, which are activated by diacylglycerol.

By using a polymerase chain reaction (PCR)-based strategy we cloned hTRPC6, which encodes a protein of the mammalian TRPC family of ion channels. hTRPC6 is a 931-amino-acid protein that shares 93% amino-acid identity with its murine homologue¹⁰. Northern blot analysis showed that hTRPC6 is expressed mainly in placenta, lung, spleen, ovary and small intestine (results not shown).

To determine the contribution of hTRPC6 to Ca²⁺ entry in response to stimulation of a receptor coupled to the G protein G_{q/11}, we microinjected nuclei of CHO-K1 cells with complementary DNA constructs encoding the H₁ histamine receptor (H₁R), a G_{q/11}-coupled receptor, and hTRPC6. The increase in [Ca²⁺]_i that occurred in cells loaded with fura-2, a fluorescent sensor of [Ca²⁺]_i, in response to stimulation with histamine was potentiated by hTRPC6 expression (Fig. 1a). Under whole-cell voltage-clamp conditions, H₁R/hTRPC6-expressing cells showed large inward currents (mean current = −2.7 ± 0.6 nA at −60 mV; n = 21 out of 22 cells studied) in

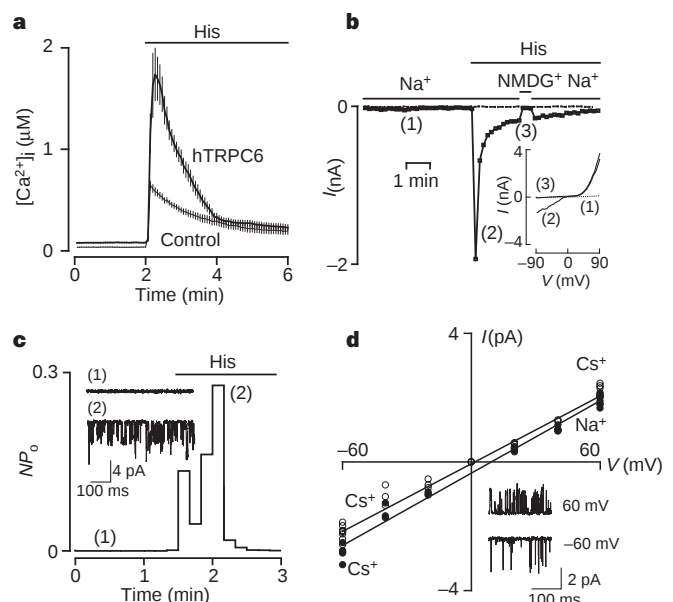


Figure 1 Functional characterization of hTRPC6. **a**, Time course of [Ca²⁺]_i in hTRPC6/H1R-expressing (n = 12) or H1R-expressing (control; n = 12) CHO-K1 cells in response to histamine. **b**, Currents in H1R/hTRPC6-expressing (filled squares) and control (dotted line) cells voltage-clamped at −60 mV. Histamine (80 μM), NMDG⁺ and Na⁺ were present in the bath (solution B1) as indicated. Inset, current traces obtained during voltage ramps from −90 to +90 mV performed at positions 1–3 of the currents. Channel activity is shown as NP_o. **c**, Stimulation of hTRPC6 channels by histamine in a cell-attached patch. Inset, sample recordings obtained at positions 1 and 2 of the graph. **d**, Current-voltage relationships in inside-out patches of hTRPC6-expressing cells assessed from mean amplitudes at various holding potentials (n = 3–7). The pipette solution P1 contained 5 μM histamine. Bath solutions were P1 (open circles) or B2 (filled circles). Insets show sample traces.

response to histamine, with a time to peak of 4.6 ± 1.3 s (Fig. 1b). The current–voltage (I – V) relation revealed dual inward and outward rectification (Fig. 1b, inset). The reversal potential of the hTRPC6 current was -3.6 ± 0.8 mV ($n = 21$). Control cells did not exhibit histamine-activated conductances ($n = 10$). Inward currents were abolished when *N*-methyl-D-glucamine (NMDG; solution B4) was the only extracellular cation ($n = 9$). An inward current was also measurable after histamine challenge when K^+ was the only extracellular cation ($n = 4$). Thus, fura-2 fluorescence and electrophysiological data indicate that hTRPC6 is a non-selective cation channel that is permeable for Ca^{2+} , Cs^+ , Na^+ and K^+ . We assessed relative ionic permeabilities by determining reversal potentials under bi-ionic conditions¹¹ with Cs^+ as the main cation in the pipette (solution P1). The relative permeabilities P_{Ca}/P_{Na} and P_{Na}/P_{Cs} were 5 and 0.7, respectively. The unitary single-channel amplitude estimated from fluctuation analysis was -1.5 ± 0.1 pA at -60 mV ($n = 8$).

In the cell-attached mode, marked single-channel activity was detected after the addition of histamine to the bath solution (Fig. 1c; $n = 8$). In inside-out patches, single-channel currents with a mean unitary current amplitude of -1.7 ± 0.1 pA were detected at a holding potential of -60 mV ($n = 18$). No such channel activity was observed in patches from control cells ($n = 9$). In studies of single channels, hTRPC6 exhibited a linear I – V relationship with a calculated slope conductance of 35 pS (duration of events >0.3 ms; reversal potential (E_{rev}) = 0.2 mV) in symmetrical 120 mM Cs^+ solutions (solution P1; Fig. 1d). Replacement of the main charge carrier in the bath solution with 120 mM Na^+ (solution B2) resulted in a shift of the reversal potential to 7.4 mV and a slope conductance of 37.5 pS. Under these conditions, the estimated permeability ratio P_{Na}/P_{Cs} was 0.75. The open probability for hTRPC6 was higher at positive (+60 mV) than at negative (-60 mV) holding potentials ($n = 4$) (Fig. 1d, inset), and the mean open time was <1 ms ($n = 17$).

hTRPC6-mediated Ca^{2+} entry could not be produced by depletion of internal stores by thapsigargin. To determine whether hTRPC6 could be activated by an increase in $[Ca^{2+}]_i$, as reported for hTRPC3 (ref. 12), we added the Ca^{2+} ionophore ionomycin (1 μ M) to hTRPC6-expressing cells. No increase in channel activity was observed ($n = 8$), but subsequent application of histamine induced an increase in single-channel activity ($n = 6$ out of 8). To examine the role of G proteins in regulation of hTRPC6, we dialysed AlF_4^- , an activator of heterotrimeric G proteins, into cells expressing hTRPC6 (30 μ M AlF_4^- in pipette solution P1). We detected NMDG-sensitive inward cation currents ($E_{rev} = -1.1 \pm 0.9$ mV) in these cells under whole-cell voltage-clamp conditions (Fig. 2a; mean current = -530 ± 93 pA, holding potential -60 mV, $n = 27$ out of 28); no cation currents could be measured in control cells ($n = 6$). AlF_4^- -stimulated currents developed within 65 ± 6 s of AlF_4^- dialysis ($n = 6$). After stimulation with AlF_4^- , the hTRPC6-expressing cells did not show further cation currents in response to histamine ($n = 3$).

A 12-min pretreatment of hTRPC6-expressing cells with 10 μ M U73122, an inhibitor of phospholipase C (PLC), abolished the histamine-induced Mn^{2+} influx in fura-2-loaded cells (Fig. 2b), whereas the biologically inactive analogue U73343 (10 μ M) had no inhibitory effect. To test for a potential role of inositol-1,4,5-trisphosphate ($InsP_3$, produced from phosphatidylinositol-4,5-bisphosphate (PtdIns(4,5) P_2) by PLC) in hTRPC6 activation, we dialysed $InsP_3$ (100 μ M in the pipette solution) into H_1R /hTRPC6-expressing cells; no currents were produced ($n = 10$) (Fig. 2c). However, these cells responded to subsequent application of histamine.

To assess the role of diacylglycerol (DAG), a second product of PLC activity, as a second messenger in stimulation of hTRPC6, we measured hTRPC6-dependent Mn^{2+} influx following application of the membrane-permeable DAG analogues 1-oleoyl-2-acetyl-*sn*-

glycerol (OAG) or 1,2-dioctanoyl-*sn*-glycerol (DOG) (100 μ M each). In hTRPC6-expressing cells, both substances elicited an instantaneous influx of Mn^{2+} , whereas control cells were unresponsive (Fig. 2d). Cellular DAG is mainly metabolized by DAG lipase¹³. We therefore treated hTRPC6-expressing cells with the DAG-lipase inhibitor RHC80267 (50 μ M), which led to an increase in quenching of fura-2 fluorescence by Mn^{2+} that was accompanied by a prolonged rise in $[Ca^{2+}]_i$ (Fig. 2e). Control cells did not respond to the inhibitor. To evaluate the contribution of protein kinase C (PKC) to the effects of DAG, we pretreated hTRPC6-expressing cells either with phorbol-12,13-didecanoate (PDD) or phorbol-12-myristoyl-13-acetate (PMA) at 2 μ M for 40 h to downregulate PKC isoforms, or with one of two PKC blockers, staurosporine (1 μ M) or bisindolylmaleimide I (500 nM), added 30 min before OAG (100 μ M) (Fig. 2f). OAG-induced Mn^{2+} influx remained unaffected by these compounds. An acute challenge of hTRPC6-expressing cells with PDD or PMA (10^{-7} to 10^{-4} M) failed to evoke rises in

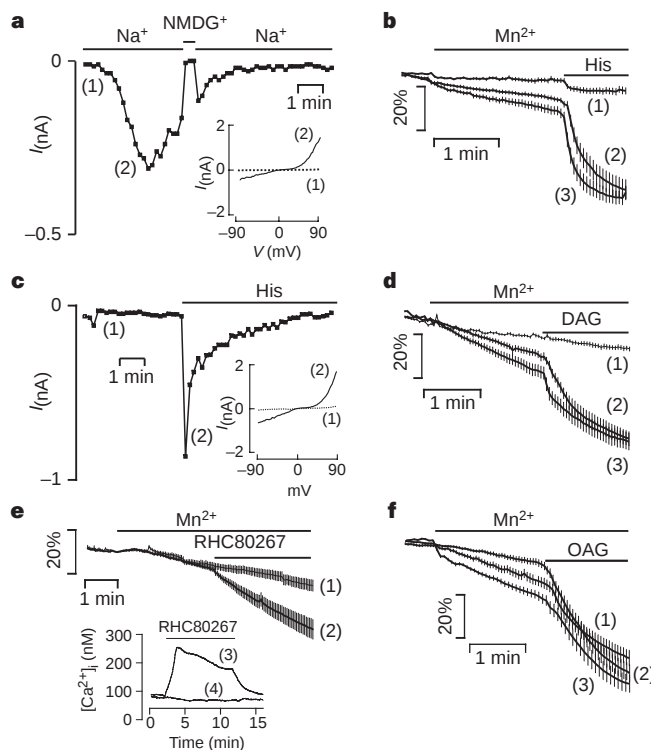


Figure 2 Regulation of hTRPC6. **a**, AlF_4^- -induced hTRPC6 currents in the presence of Na^+ and NMDG $^+$ at -60 mV in the whole-cell mode. Inset, current traces obtained during voltage ramps from -90 to $+90$ mV at positions 1 and 2 of the main current trace. **b**, Agonist-induced Mn^{2+} quenching in H_1R /hTRPC6-expressing cells. Cells were pretreated for 12 min with 1, 10 μ M U73122 ($n = 18$); 2, 0.2% dimethyl sulphoxide ($n = 12$); or 3, 10 μ M U73343 ($n = 18$). Percentages shown are per cent quenching of fura-2 fluorescence by Mn^{2+} . **c**, Effect of dialysis of $InsP_3$ (100 μ M) on histamine-induced hTRPC6 currents. H_1R /hTRPC6-expressing cells were voltage-clamped at -60 mV. Inset, current traces obtained during voltage ramps from -90 to $+90$ mV obtained at positions 1 and 2 of the main current trace. Solution P1 was in the pipette and solution B1 was in the bath. **d**, Mn^{2+} influx evoked by membrane-permeable DAG analogues. 1, H_1R -expressing cells ($n = 8$) stimulated with 100 μ M OAG; H_1R /hTRPC6-expressing cells exposed to 2, 100 μ M OAG ($n = 15$) or 3, 100 μ M DOG ($n = 15$ cells). **e**, hTRPC6 activation by a DAG-lipase inhibitor. RHC80267 (50 μ M) was applied to 1, control ($n = 9$) and 2, hTRPC6-expressing ($n = 14$) cells; Mn^{2+} quenching of fura-2 fluorescence was recorded. Inset, time courses of $[Ca^{2+}]_i$ in 3, hTRPC6-expressing and 4, control cells. **f**, Effect of PKC inhibition on OAG-induced Mn^{2+} quenching. hTRPC6-expressing cells were 1, not treated ($n = 18$), 2, pre-incubated for 40 h with 2 μ M PDD ($n = 14$), or 3, exposed to 1 μ M staurosporine ($n = 13$) before OAG challenge.

[Ca²⁺]_i. In whole-cell experiments, pre-incubation with staurosporine or calphostin C (1 μM), a competitive inhibitor of binding of phorbol esters and DAG to PKC isoforms, did not inhibit hTRPC6 activation. Together, these results indicate a PKC-independent effect of diacylglycerols on hTRPC6.

To determine whether hTRPC6 could be activated in a membrane-delimited fashion, we studied the effects of various DAG analogues in inside-out patches. Addition of the membrane-permeable OAG (100 μM) as well as the impermeable, endogenous DAG analogues 1-stearoyl-2-arachidonoyl-*sn*-glycerol (SAG, 10 μM) and 1-stearoyl-2-linoleoyl-*sn*-glycerol (SLG, 100 μM) to the bath evoked a marked increase in channel activity (Fig. 3a–c), with SAG (*n* = 18 out of 22) being the most effective (mean increase in *NP*_o, the product of the number of channels (*N*) and the channel open probability (*P*_o), was ~100-fold). SAG's effect was less at 1 μM (mean increase in *NP*_o ~6-fold; *n* = 3 out of 4), and no effect of SAG was detected at 100 nM (*n* = 3 out of 3).

Application of impermeable DAG analogues at the above concentrations to the incubation medium of fura-2-loaded intact hTRPC6-expressing cells did not result in increases in [Ca²⁺]_i (*n* = 6). Exposure of inside-out patches of control cells to 10 μM SAG (*n* = 11) did not reveal channel activation. Patches obtained from cells expressing the rat vanilloid receptor subtype 1 (rVR1)¹⁴ failed to respond to SAG but did respond to 10 μM capsaicin with

marked channel activity (Fig. 3d). Conversely, capsaicin did not evoke currents in hTRPC6-containing membrane patches (*n* = 5). Addition of 10 μM InsP₃ to inside-out patches isolated from hTRPC6-expressing cells confirmed that this second messenger cannot activate hTRPC6, whereas subsequent SAG application resulted in sustained hTRPC6 activity (*n* = 4) (Fig. 3e). We tested the ability of PtdIns(4,5)P₂ (10 μM), a direct activator of inward rectifier K⁺ channels¹⁵, to stimulate hTRPC6 (Fig. 3f). PtdIns(4,5)P₂ had no effect on hTRPC6 activation (*n* = 9) but the channel did respond to subsequent application of SAG (*n* = 4 out of 6). PMA (1 μM) had no effect on hTRPC6-containing inside-out patches (*n* = 5), which subsequently reacted to SAG (*n* = 3 out of 4). The addition of two monoacylglycerols, 1-oleoyl-glycerol (*n* = 3) and 2-oleoyl-glycerol (*n* = 3), at 100 μM did not increase hTRPC6 single-channel activity in inside-out patches.

Several mammalian members of the TRPC family, namely hTRPC3 (refs 12, 16), mTRPC5 (ref. 17) and hTRPC6 (ref. 10), are activated after stimulation of G_{q/11}-coupled receptors, yet the

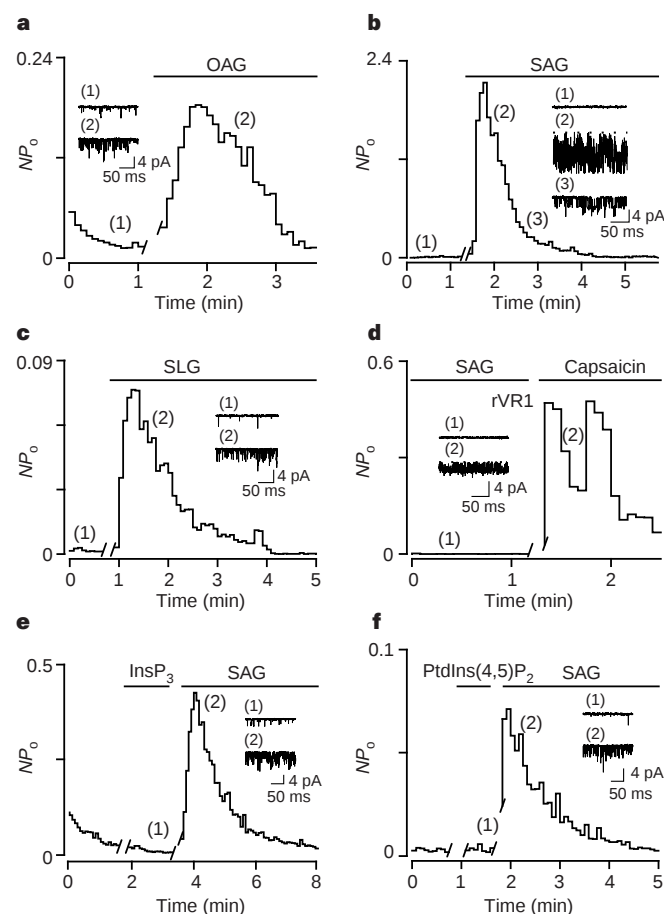


Figure 3 Effect of DAG derivatives on hTRPC6 and rVR1 channels in inside-out patches. Cells were injected with cDNAs encoding hTRPC6 (**a**, **b**, **c**, **e**, **f**) or rVR1 (**d**). Channel activity is shown as *NP*_o. **a**, OAG (100 μM); **b**, SAG (10 μM); **c**, SLG (100 μM); **d**, SAG and capsaicin (10 μM); **e**, InsP₃ and SAG (10 μM); and **f**, PtdIns(4,5)P₂ and SAG (10 μM) were added to the bath (solution B2); solution B3 was used in the pipette. The holding potential was -60 mV. Inset show sample traces collected at positions 1 and 2 (or 1–3 for SAG) of each graph.

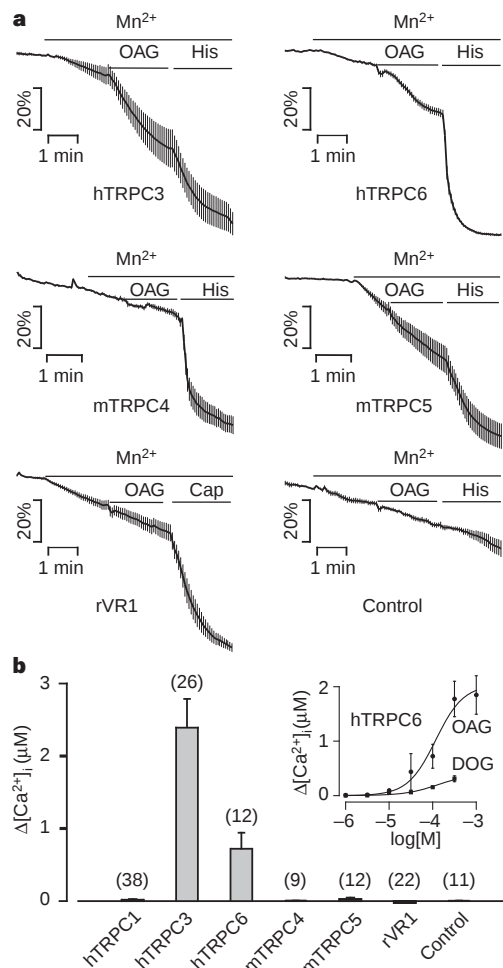


Figure 4 Selective action of DAG derivatives on a subfamily of TRPC channels. **a**, Effect of OAG on TRPC channels. Expression constructs containing the channels indicated or pcDNA3 (control vector) were microinjected together with H₁R cDNA. Cells were stimulated with 100 μM OAG and then 100 μM histamine, and Mn²⁺ quenching of fura-2 fluorescence was monitored. **b**, OAG-induced activation of TRPC channels. CHO cells microinjected with expression constructs or with pcDNA3 (control) were used to monitor the increase in [Ca²⁺]_i (Δ[Ca²⁺]_i) evoked by 100 μM OAG. The numbers of cells studied are indicated in parentheses. Inset, dose-response relationship for OAG and DOG in hTRPC6-expressing cells (*n* = 4–18 for each data point).

mechanism of channel activation is poorly understood. hTRPC1 and bTRPC4 have been classified as store-operated Ca^{2+} -entry channels^{18,19}. An alignment of the amino-acid sequences of the TRPC family members reveals a significantly higher identity between TRPCs 3 and 6, as well as between TRPCs 4 (ref. 20) and 5, in comparison with other TRPC proteins, indicating that these structural similarities may also have a functional correlate. We therefore studied OAG- and histamine-induced Mn^{2+} influx in cells expressing H_1R together with hTRPC3, hTRPC6, mTRPC4 or mTRPC5. Cells expressing rVR1 or injected with H_1R cDNA only served as controls (Fig. 4a). OAG (100 μM) stimulated Mn^{2+} influx in hTRPC3- and hTRPC6-expressing cells only, and not in cells expressing mTRPC4 or mTRPC5. Histamine application, however, gave rise to profound Mn^{2+} quenches in cells expressing TRPCs 3, 6, 4 or 5, showing that all of these channels are activated after stimulation of a $\text{G}_{q/11}$ -coupled receptor. Neither rVR1-expressing nor control-injected cells responded to OAG challenge by increasing Mn^{2+} influx. Only cells expressing hTRPCs 3 and 6 showed an OAG-induced increase in $[\text{Ca}^{2+}]_i$; hTRPC1, mTRPC4, mTRPC5, rVR1 and control-injected cells remained unresponsive (Fig. 4b). The effect of OAG on hTRPC6 was concentration-dependent, with an effector concentration for half-maximal response of 117 μM (Fig. 4b, inset). Control cells never exhibited OAG- or DOG-induced Mn^{2+} influx.

There are reasons to doubt that light-activated ion channels in *Drosophila* photoreceptors are gated exclusively by store depletion (summarized in refs 21, 22). Here we have delineated an alternative pathway for activation of such channels and have shown that hTRPC6 and hTRPC3 are activated by diacylglycerols in a store-depletion-independent, membrane-delimited manner. Our results may explain the rapid onset of initial Ca^{2+} influx in response to stimulation of $\text{G}_{q/11}$ -coupled receptors observed in several experimental systems²³, as well as the requirement of PLC for the activation of hTRPC3 (ref. 16) and hTRPC6. Diacylglycerol-mediated activation of non-selective cation currents has been observed in rabbit portal vein myocytes²⁴. So far, however, hTRPC6 and hTRPC3 represent, to our knowledge, the first molecularly defined ion channels that are known to be activated by diacylglycerols independently of protein kinase C. □

Methods

Molecular cloning. hTRPC6 was cloned from human testis and placental total RNA by applying a PCR-based approach, and was subcloned into the expression vector pcDNA3 (Stratagene). VR1 (ref. 14), mTRPC4 (ref. 20), and mTRPC5 (ref. 17) cDNAs were isolated from rat dorsal root ganglia or mouse brain total RNA by using a similar strategy.

Cell culture and microinjection. CHO-K1 cells were seeded onto glass coverslips and cultured for 24 h. Intracellular microinjection was done as described¹². Injection mixtures contained 5 mM Tris/HCl, pH 7.5, various combinations of 0.5 $\mu\text{g } \mu\text{L}^{-1}$ channel cDNAs in pcDNA3, 0.005 $\mu\text{g } \mu\text{L}^{-1}$ of H_1R cDNAs in pCMV, and 0.002 $\mu\text{g } \mu\text{L}^{-1}$ enhanced green-fluorescent protein cDNA in pEGFP-C1. Control cells were injected with 'empty' pcDNA3 vector.

Assessment of $[\text{Ca}^{2+}]_i$ and Mn^{2+} influx. Cells were loaded with fura-2-AM and superfused with buffer A containing (in mM): 120 NaCl, 6 KCl, 1.25 MgCl_2 , 5.5 glucose and 10 HEPES, pH 7.4, at a constant flow rate of 5 ml min^{-1} . Fura-2 fluorescence was recorded over manually defined regions of interest, with excitation alternating between 359 and 380 nm at 4-s intervals, using a digital fluorescence imaging system (TILL Photonics). To determine Mn^{2+} quenching, we added 200 μM MnCl_2 2 min before agonist stimulation. The decrease in fluorescence was monitored at the isosbestic wavelength (359 nm). Background fluorescence was determined by supplementing buffer A with 10 μM Triton-X100 and 10 mM MnCl_2 . $[\text{Ca}^{2+}]_i$ was determined as described²⁵. All data are presented as means $\pm$ s.e.m.

Electrophysiology. The patch-clamp technique²⁶ in whole-cell, cell-attached and inside-out mode was used to study hTRPC6, as well as rVR1 expressed in CHO-K1 cells. Solution B1 contained (in mM): 140 sodium isothionate, 5 potassium gluconate, 1.8 calcium gluconate, 1 magnesium gluconate, 10

glucose and 10 HEPES; solution B2 contained: 120 sodium isothionate, 5.87 calcium gluconate, 1 magnesium gluconate, 10 EGTA, 10 glucose and 10 HEPES; solution B3 contained: 120 CsCl, 1.8 calcium gluconate, 1 magnesium gluconate, 10 glucose and 10 HEPES; solution B4 contained: 140 NMDG isothionate, 5 EGTA, 10 glucose and 10 HEPES; solution B5 contained: 120 sodium isothionate, 1 EGTA, 10 glucose and 10 HEPES; solution B6 contained: 10 calcium gluconate, 130 NMDG isothionate, 10 glucose and 10 HEPES; solution B7 contained: 120 CsCl, 1 EGTA, 10 glucose and 10 HEPES; pipette solution P1 contained: 120 CsCl, 5.87 calcium gluconate, 1 magnesium gluconate, 10 EGTA and 10 HEPES. Solutions were buffered to pH 7.4. The osmolality was adjusted to 290–310 mosM with mannitol. An agar bridge served as the electrical connection between the bath and the signal ground. In whole-cell experiments, the access resistance was less than 10 M Ω and series resistance compensation was set to 65–85%. For fluctuation analysis²⁷, no series resistance compensation was used. Reversal potentials (E) of currents were determined from currents recorded during voltage ramps. Measurements were corrected for liquid-junction potentials. Relative ion permeabilities for monovalent cations were calculated as described¹¹. The $P_{\text{Ca}}/P_{\text{Na}}$ permeability ratio was calculated according to the equation $P_{\text{Ca}}/P_{\text{Na}} = \{[\text{Na}^+]_o \times \exp(-FE_{\text{Na}}/RT) \exp(FE_{\text{Ca}}/RT) / (\exp(FE_{\text{Ca}}/RT) + 1)\} / (4[\text{Ca}^{2+}]_o)$, where R , T and F are the gas constant, absolute temperature and Faraday's constant, respectively. In these experiments, bath solutions contained only one of the following cations: Na^+ (solution B5, $E_{\text{Na}} = -10.2 \pm 1.5$ mV, $n = 15$), Ca^{2+} (solution B6, $E_{\text{Ca}} = -13.4 \pm 4.15$ mV, $n = 12$), or Cs^+ (solution B7, $E_{\text{Cs}} = 1.6 \pm 2.1$ mV, $n = 3$). Analysis was performed with pClamp 6 software (Axon Instruments). Channel activity is expressed as N_{Po} , calculated for consecutive 5-s intervals. In whole-cell experiments, data were filtered at 1 kHz; in single-channel experiments, data were filtered at 2.5 kHz. Bath solutions containing SAG, SLG, DOG, OAG and PtdIns(4,5) P_2 were sonicated for 5 min before use. All experiments were performed at room temperature (21–26 °C). Data are presented as means $\pm$ s.e.m.

Received 31 July; accepted 17 November 1998.

- Clapham, D. E. Calcium signaling. *Cell* **80**, 259–268 (1995).
- Hoth, M. & Penner, R. Depletion of intracellular calcium stores activates a calcium current in mast cells. *Nature* **355**, 353–356 (1992).
- Zweifach, A. & Lewis, R. S. Mitogen-regulated Ca^{2+} current of T lymphocytes is activated by depletion of intracellular Ca^{2+} stores. *Proc. Natl Acad. Sci. USA* **90**, 6296–6299 (1993).
- Berridge, M. J. Capacitative calcium entry. *Biochem. J.* **312**, 1–11 (1995).
- Putney, J. W. Jr A model for receptor-regulated calcium entry. *Cell Calcium* **7**, 1–12 (1986).
- Fasolato, C., Hoth, M., Matthews, G. & Penner, R. Ca^{2+} and Mn^{2+} influx through receptor-mediated activation of nonspecific cation channels in mast cells. *Proc. Natl Acad. Sci. USA* **90**, 3068–3072 (1993).
- Hardie, R. C. & Minke, B. Novel Ca^{2+} channels underlying transduction in *Drosophila* photoreceptors: implications for phosphoinositide-mediated Ca^{2+} mobilization. *Trends Neurosci.* **16**, 371–376 (1993).
- Zhu, X. *et al.* A novel mammalian gene family essential for agonist-activated capacitative Ca^{2+} entry. *Cell* **85**, 661–671 (1996).
- Montell, C. New light on TRP and TRPL. *Mol. Pharmacol.* **52**, 755–763 (1997).
- Boulay, G. *et al.* Cloning and expression of a novel mammalian homolog of *Drosophila* transient receptor potential (Trp) involved in calcium entry secondary to activation of receptors coupled by the G_q class of G protein. *J. Biol. Chem.* **272**, 29672–29680 (1997).
- Hille, B. *Ionic Channels of Excitable Membranes* (Sinauer, Sunderland, Massachusetts, 1992).
- Zitt, C. *et al.* Expression of TRPC3 in chinese hamster ovary cells results in calcium-activated cation currents not related to store depletion. *J. Cell Biol.* **138**, 1333–1341 (1997).
- Migas, I. & Severson, D. L. Diacylglycerols derived from membrane phospholipids are metabolized by lipases in A10 smooth muscle cells. *Am. J. Physiol.* **271**, C1194–C1202 (1996).
- Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
- Huang, C.-L., Feng, S. & Hilgemann, D. W. Direct activation of inward rectifier potassium channels by PIP_2 and its stabilization by G $\beta\gamma$. *Nature* **391**, 803–806 (1998).
- Zhu, X., Jiang, M. & Birnbaumer, L. Receptor-activated Ca^{2+} influx via human Trp3 stably expressed in human embryonic kidney (HEK) 293 cells. Evidence for a non-capacitative Ca^{2+} entry. *J. Biol. Chem.* **273**, 133–142 (1998).
- Okada, T. *et al.* Molecular cloning and functional characterization of a novel receptor-activated TRP Ca^{2+} channel from mouse brain. *J. Biol. Chem.* **273**, 10279–10287 (1998).
- Zitt, C. *et al.* Cloning and functional expression of a human Ca^{2+} -permeable cation channel activated by calcium store depletion. *Neuron* **16**, 1189–1196 (1996).
- Philipp, S. *et al.* A mammalian capacitative calcium entry channel homologous to *Drosophila* TRP and TRPL. *EMBO J.* **15**, 6166–6171 (1996).
- Mori, Y. *et al.* Differential distribution of TRP Ca^{2+} channel isoforms in mouse brain. *Neuroreport* **9**, 507–515 (1998).
- Scott, K. & Zuker, C. TRP, TRPL and trouble in photoreceptor cells. *Curr. Opin. Neurobiol.* **8**, 383–388 (1998).
- Montell, C. TRP trapped in fly signaling web. *Curr. Opin. Neurobiol.* **8**, 389–397 (1998).
- Birnbaumer, L. *et al.* On the molecular basis and regulation of cellular capacitative calcium entry: roles for Trp proteins. *Proc. Natl Acad. Sci. USA* **93**, 15195–15202 (1996).
- Helliwell, R. M. & Large, W. A. α_1 -Adrenoceptor activation of a non-selective cation current in rabbit portal vein by 1,2-diacyl-sn-glycerol. *J. Physiol.* **499**, 417–428 (1997).
- Dippel, E., Kalkbrenner, F., Wittig, B. & Schultz, G. A heterotrimeric G protein complex couples the muscarinic m1 receptor to phospholipase C- β . *Proc. Natl Acad. Sci. USA* **93**, 1391–1396 (1996).
- Hamill, O. P., Marty, A., Neher, E., Sakmann, B. & Sigworth, F. J. Improved patch-clamp techniques

- for high-resolution current recording from cells and cell-free membrane patches. *Pflügers Arch.* **391**, 85–100 (1981).
27. Neher, E. & Stevens, C. F. Conductance fluctuations and ionic pores in membranes. *Annu. Rev. Biophys. Bioeng.* **6**, 345–381 (1977).

Acknowledgements. We thank N. Albrecht for technical help with cell cultures, J.-C. Schwartz for providing the guinea pig H₁ histamine receptor cDNA, and T. Plant for critically reading the manuscript. This work was supported by grants from the Deutsche Forschungsgemeinschaft.

Correspondence and requests for materials should be addressed to G.S. (e-mail: gschultz@zedat.fu-berlin.de). The hTRPC6 cDNA sequence has been deposited at Genbank under accession number AF080394.

ICOS is an inducible T-cell co-stimulator structurally and functionally related to CD28

Andreas Hutloff*, Anna M. Dittrich*†, Katja C. Beier*†, Barbara Eljaschewitsch*, Regine Kraft‡, Ionnis Anagnostopoulos§ & Richard A. Krocze*

* Molecular Immunology, Robert Koch-Institut, Nordufer 20, 13353 Berlin, Germany

‡ Department of Protein Chemistry, Max-Delbrück-Centrum, Robert-Rössle-Str. 10, 13122 Berlin, Germany

§ Institute of Pathology, Klinikum Benjamin Franklin, Freie Universität Berlin, Hindenburgdamm 30, 12200 Berlin, Germany

† These authors contributed equally to this work

The T-cell-specific cell-surface receptors CD28 and CTLA-4 are important regulators of the immune system. CD28 potentially enhances those T-cell functions that are essential for an effective antigen-specific immune response^{1–5}, and the homologous CTLA-4 counterbalances the CD28-mediated signals and thus prevents an otherwise fatal overstimulation of the lymphoid system^{6–9}. Here we report the identification of a third member of this family of molecules, inducible co-stimulator (ICOS), which is a homodimeric protein of relative molecular mass 55,000–60,000 (*M_r*, 55K–60K). Matching CD28 in potency, ICOS enhances all basic T-cell responses to a foreign antigen, namely proliferation, secretion of lymphokines, upregulation of molecules that mediate cell–cell interaction, and effective help for antibody secretion by B cells. Unlike the constitutively expressed CD28, ICOS has to be *de novo* induced on the T-cell surface, does not upregulate the production of interleukin-2, but superinduces the synthesis of interleukin-10, a B-cell-differentiation factor. *In vivo*, ICOS is highly expressed on tonsillar T cells, which are closely associated with B cells in the apical light zone of germinal centres, the site of terminal B-cell maturation. Our results indicate that ICOS is another major regulator of the adaptive immune system.

We identified ICOS by generating monoclonal antibodies against activated human T cells. The ICOS-specific monoclonal antibody F44 did not react with resting human peripheral-blood T cells, but stained CD4⁺ and CD8⁺ T lymphocytes that had been activated by stimulation of the T-cell antigen receptor (TCR) complex (Fig. 1a). No signal was detected on resting or appropriately activated B cells, monocytes, natural killer cells, granulocytes, dendritic cells or platelets (data not shown). Immunoprecipitations using monoclonal antibody F44 defined the ICOS antigen as a disulphide-linked dimer, with an apparent relative molecular mass of 55K–60K, composed of a 27K and a 29K chain (Fig. 1b). We enriched ICOS protein from lysates of the ICOS-expressing T-cell line MOLT-4V by large-scale affinity chromatography using F44, and further purified the protein by two-dimensional preparative SDS–polyacrylamide gel electrophoresis (PAGE) (Fig. 1b). The 27K and 29K ICOS-protein species yielded identical peptide sequences, indicating that ICOS may be expressed on the cell surface as a homodimeric

protein, with the two chains differing only in their post-translational modification; this assumption was later confirmed by transfection experiments (data not shown).

Using the peptide sequence, we cloned full-length ICOS complementary DNA (2,641 base pairs) from a MOLT-4V cDNA library. Northern analysis revealed a single ICOS messenger RNA species of ~2.8 kilobases in length in activated human T cells (Fig. 1c). The open reading frame of ICOS mRNA encodes a new protein of 199 amino acids with a predicted relative molecular mass of 22.6K. The ICOS amino-acid sequence shares 24% (17%) identity and 39% (39%) similarity with CD28 (Fig. 1d) (and CTLA-4 (ref. 10)). The predicted mature ICOS is a type-I transmembrane molecule (type I transmembrane proteins have their carboxy termini in the cytosol) that consists of a single immunoglobulin (Ig)V-like domain (stabilized by conserved cysteine residues at positions 42 and 109; Fig. 1d), a transmembrane region of ~23 amino acids, and a cytoplasmic tail of 35 amino acids, and it shows a close structural resemblance to CD28 (Fig. 1d) and CTLA-4 (ref. 10). The cysteine residue located at position 141 of CD28, also found in CTLA-4, is apparently involved in forming the disulphide bridge between the homodimeric chains of these proteins, and is also found in ICOS (position 136). The motif MYPPPY (positions 117–122 in CD28), required in its intact form for the binding of CD28 and CTLA-4 to their counter-receptors B7-1 and B7-2 (refs 11, 12), is not conserved in ICOS, indicating that ICOS may interact with a different receptor.

In a series of *in vitro* experiments, we compared ICOS to the main co-stimulatory effects of CD28 directly. We tested the ICOS-specific monoclonal antibody F44 in parallel with monoclonal antibody 9.3, one of the most potent CD28-specific reagents available^{1,13}. When human peripheral-blood CD4⁺ T cells were suboptimally activated through the CD3 complex of the TCR, co-stimulation of T-cell proliferation by ICOS was comparable in potency to signalling by CD28 (shaded area in upper panel of Fig. 2a). This finding was remarkable, given that ICOS was present on at most 30–40% of the CD4⁺ T cells under these experimental conditions (shaded area in lower panel of Fig. 2a), whereas CD28 continued to be expressed on nearly all CD4⁺ T cells. At higher dilutions of the CD3-specific monoclonal antibody OKT 3, the co-stimulatory effect of monoclonal antibody F44 could no longer be achieved, because the signal through the TCR complex was insufficient to induce the expression of the ICOS molecule on the T-cell surface (Fig. 2a, lower panel).

Only a small amount of most lymphokines is secreted from primary CD4⁺ T cells after optimal triggering of the CD3/TCR complex alone¹⁴. One characteristic of co-stimulation of CD4⁺ T cells by CD28 is the enhancement of cytokine secretion, in particular the ‘superinduction’ of interleukin (IL)-2 (Fig. 2b and refs 1, 3, 4, 14). Surprisingly, co-stimulation by ICOS failed to significantly upregulate IL-2 at any time point (Fig. 2b). However, co-stimulation by ICOS upregulated the secretion of IL-4, IL-5, interferon- γ , tumour-necrosis factor- α and granulocyte/macrophage-colony-stimulating factor to 50–70% of the levels achieved with co-stimulation by CD28 (Fig. 2b and data not shown). ICOS and CD28 had markedly different co-stimulatory effects on the secretion of IL-10. In contrast to the modest stimulatory capacity of CD28, co-stimulation by ICOS routinely induced several-fold higher secretion of IL-10 (Fig. 2b); northern blot analysis confirmed the upregulation of expression of IL-10 (data not shown). As the proportion of ICOS⁺ T cells in the lymphokine-co-stimulation system did not exceed 50% (Fig. 2a, lower panel), on a per-cell basis ICOS matched the ability of CD28 to amplify the secretion of many lymphokines (except for IL-2), and was superior to CD28 in co-inducing IL-10.

T cells communicate with other cells of the immune system by *de novo* expression of cell-surface molecules, as well as by generating lymphokines. Co-stimulation of T cells through ICOS or CD28 markedly upregulated expression of CD154 (also known as CD40 ligand or TRAP), a molecule required for the crosstalk of T cells with

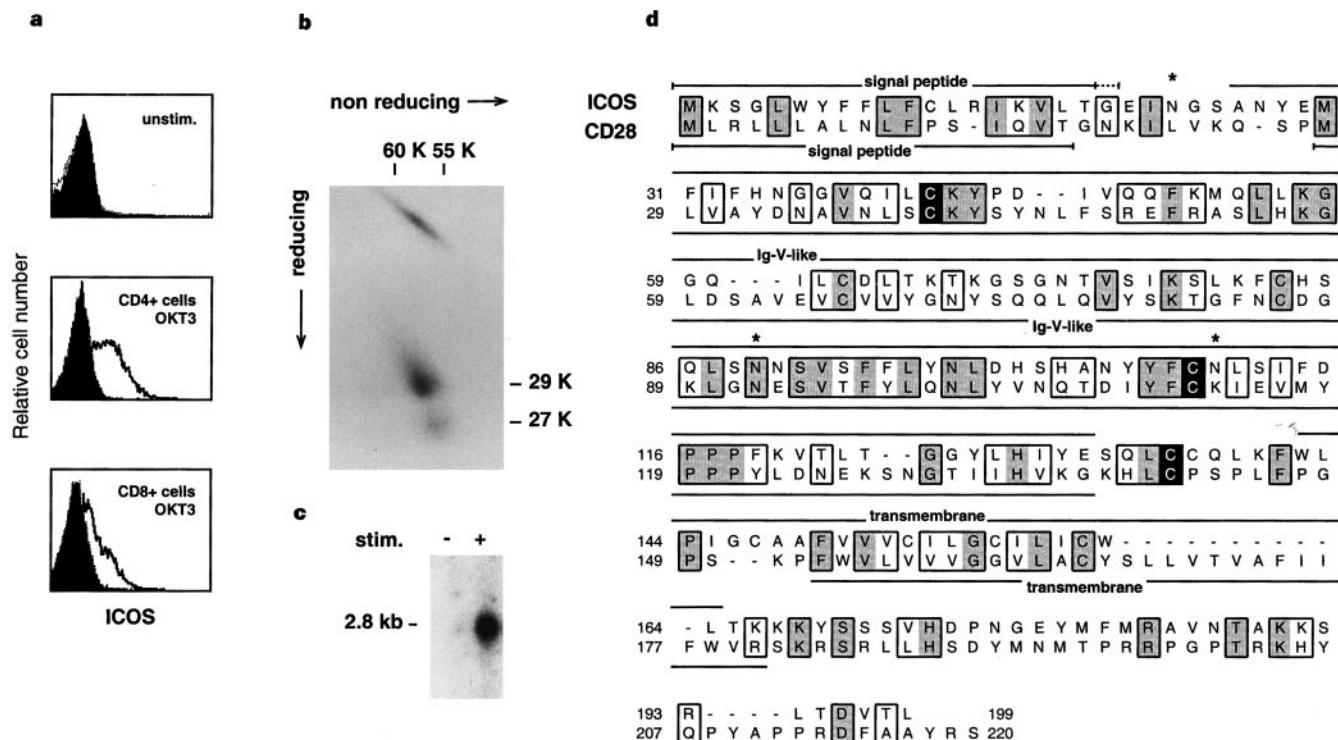


Figure 1 Identification, purification and cloning of ICOS. **a**, Expression of ICOS on human T cells after 36 h of stimulation. Peripheral-blood CD4⁺ or CD8⁺ T cells were left untreated or were stimulated with the solid-phase-bound anti-CD3 monoclonal antibody OKT 3 (1:1,000 dilution of ascites), and were analysed by flow cytometry using the fluorescein isothiocyanate (FITC)-labelled monoclonal antibody F44. At 14 h after stimulation, ICOS could not yet be detected on CD8⁺ T cells, whereas CD4⁺ T cells expressed levels of ICOS that were similar to those expressed after 36 h (data not shown). The background signal obtained with the isotype-control monoclonal antibody MOPC-21 (Sigma) is shown in black. Stimulation by phorbol-12-myristate-13-acetate (PMA) and ionomycin led to the expression of ICOS on most CD4⁺ and CD8⁺ T cells (data not shown). **b**, Structure of the homodimeric ICOS protein. ICOS protein was immunoprecipitated from surface-iodinated MOLT-4V cells with monoclonal antibody F44 and separated by two-dimensional (non-reducing/reducing) SDS-PAGE. Numbers at the top and right side of the gel are *M_r* values. The 55K–60K protein species on the diagonal

corresponds to residual dimeric ICOS, which was not reduced by the in-gel reducing procedure required for the two-dimensional SDS-PAGE (a full reduction of this species was routinely observed in one-dimensional SDS-PAGE; data not shown). Identical data were obtained with iodinated activated primary T cells (data not shown). **c**, ICOS mRNA expression. Amounts of ICOS mRNA were determined by northern blot analysis of total RNA (2.5 µg per lane) obtained from human peripheral-blood CD4⁺ T cells activated with PMA (20 ng ml⁻¹) and ionomycin (200 ng ml⁻¹) for 24 h. **d**, Amino-acid-sequence alignment of ICOS and CD28 (ref. 28), obtained using the Clustal-W algorithm with BLOSUM 30 matrix (MacVector, Oxford Molecular Group). Identical amino-acid residues are shaded; conserved residues are boxed; potential *N*-glycosylation sites are indicated by asterisks. Predicted signal peptides (the exact cleavage site for ICOS has not been determined experimentally), IgV-like domains²⁹ and predicted transmembrane regions are indicated. Cysteine residues that may be important in maintaining three-dimensional structure are shown in black.

B cells¹⁵ (Fig. 2c), and also other early T-cell-activation antigens such as CD69, CD25 or CD71 (data not shown). Next, we tested whether the ICOS-induced expression of a variety of lymphokines and cell-surface receptors was sufficient for the complex communication of T cells with tonsillar B lymphocytes. Unstimulated T cells, or T cells that had only been preactivated through the TCR, could not induce immunoglobulin synthesis by B cells (Fig. 2d). Only co-stimulation through ICOS or CD28 enabled T cells to provide effective help for B cells to synthesize IgM and IgG (Fig. 2d). On a per-cell basis, co-stimulation through ICOS was as effective as co-stimulation through CD28 in upregulating the expression of cell-surface receptors by T cells or IgG synthesis by B cells.

To determine the function of ICOS *in vivo*, we stained tissue sections of various lymphoid organs with monoclonal antibody F44. In the adult thymus and spleen, ICOS expression was almost completely confined to the few small germinal centres that are present in these tissues (data not shown). In tonsils, a substantial number of cells stained positively for ICOS and we therefore characterized tonsillar cells extensively by flow cytometry. Of these tonsillar T cells, 50–70% expressed high levels of ICOS (Fig. 3a); all other tonsillar cell populations did not express ICOS (data not shown). The expression of ICOS was associated with the

presence of CD28 on the T-cell surface (Fig. 3a). Nearly all ICOS⁺ T cells also expressed CD45RO (Fig. 3a), indicating that T cells bearing high levels of ICOS may be in a late phase of activation; ICOS⁺ T cells had the phenotype of resting cells (data not shown). Finally, almost all T cells positive for CD57, a marker for a subpopulation of T cells that are located in germinal centres of lymphoid tissues^{16,17}, co-expressed ICOS (Fig. 3a). Detailed immunohistological studies showed that ICOS is predominantly expressed on germinal-centre T cells located in the apical part of the germinal-centre light zone (Fig. 3b), a site in which T cells are known to induce terminal differentiation of B cells into antibody-secreting plasma cells or memory cells^{16,17}. We also identified a smaller but significant proportion of ICOS⁺ cells in the T-cell zone surrounding germinal centres (Fig. 3b).

On the basis of its structure, its T-cell-restricted expression and its function, we conclude that ICOS is closely related to CD28, an important molecule in the immune system^{2–5,16,17}. However, despite their overall similarity, ICOS and CD28 differ profoundly in several aspects. In humans, CD28 is constitutively expressed at high levels on almost all CD4⁺ and on 50% of CD8⁺ T cells, whereas expression of ICOS must be induced *de novo*. Both CD28 and ICOS co-induce the synthesis of several lymphokines, but only CD28 superinduces

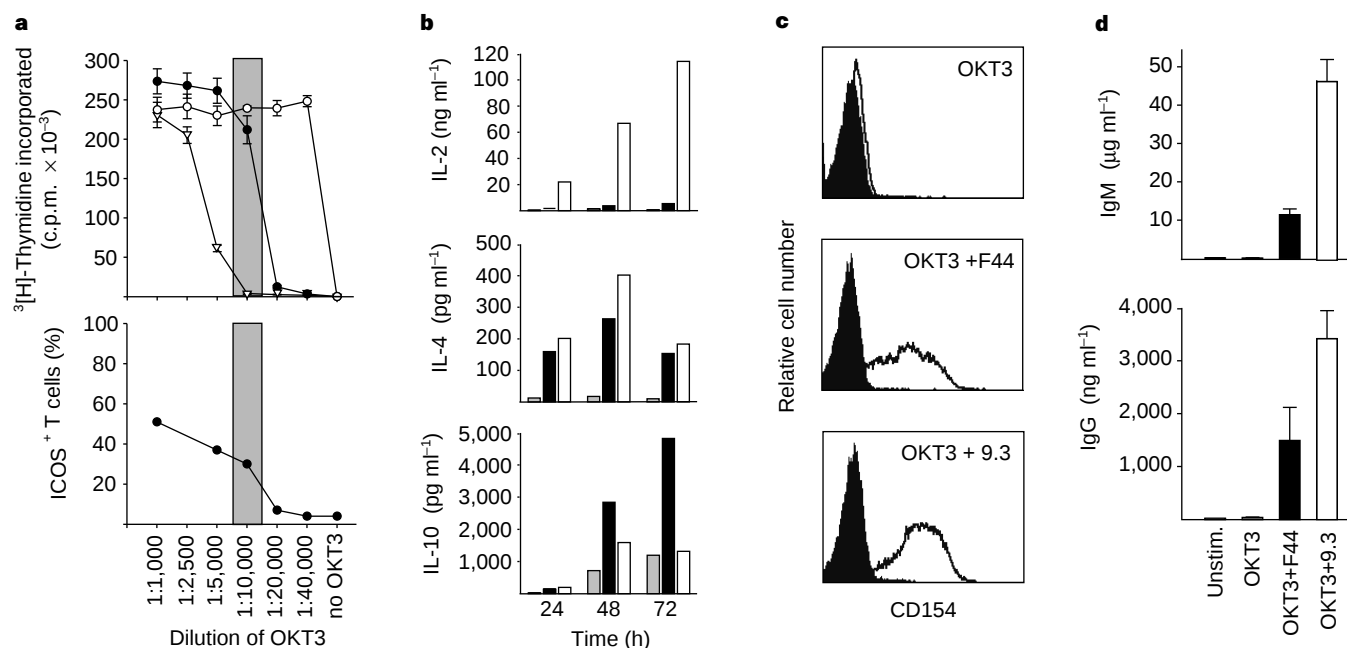


Figure 2 Co-stimulatory functions of ICOS. **a**, Co-stimulation of T-cell proliferation. Peripheral-blood CD4 $^{+}$ T cells were cultured in the presence of monoclonal antibody OKT 3 (dilutions of ascites), and monoclonal antibodies 9.3 (open circles), F44 (filled circles), or isotype-control MOPC-21 (triangles) (see Methods). Top, proliferation of T cells was determined by addition of 1 μCi [^3H]thymidine for the last 16 h of the 72 h culture period. The shaded area shows the experimental point at which maximal co-stimulation with F44 was observed. In ten experiments, co-stimulation through ICOS enhanced proliferation of T cells by 60- to 80-fold (range 14- to 135-fold) over the background stimulation observed with OKT 3 alone. Bottom, the percentage of T cells induced to express ICOS at the various dilutions of OKT 3 ascites in cultures with co-crosslinked F44 at 24 h after stimulation, the time point of maximal ICOS expression. **b**, Co-stimulation of lymphokine secretion by primary T cells. Peripheral-blood CD4 $^{+}$ cells were stimulated by optimal CD3-crosslinking (1:1,000 dilution of OKT 3 ascites) in the presence of co-stimulating monoclonal antibody 9.3 (white columns) or F44 (black columns), or in the presence of the control monoclonal antibody MOPC-21

(stippled columns). Supernatants were collected at the indicated times after stimulation and assayed for various cytokines using enzyme-linked immunosorbent assays (ELISAs) (Biosource). Results of a representative experiment out of six are shown. **c**, Upregulation of the cell-surface molecule CD154 (also known as CD40 ligand and TRAP). Peripheral blood CD4 $^{+}$ T cells were suboptimally stimulated (1:5,000 dilution of OKT 3 ascites) in the presence of control monoclonal antibody MOPC-21 or monoclonal antibodies F44 or 9.3, and CD154 expression was determined by flow cytometry with an FITC-labelled antibody, TRAP1 (ref. 15). **d**, Effect of T-cell co-stimulation through ICOS on immunoglobulin synthesis by tonsillar B cells. Peripheral-blood CD4 $^{+}$ T cells were cultured with tonsillar B cells in microtitre plates precoated with monoclonal antibody OKT 3 (1:5,000 dilution of ascites) and MOPC-21, F44 or 9.3. IgM and IgG levels in the culture supernatants were determined by ELISA on day 8. Error bars indicate the s.d. from quadruplicate cultures. Results from a representative experiment out of six are shown.

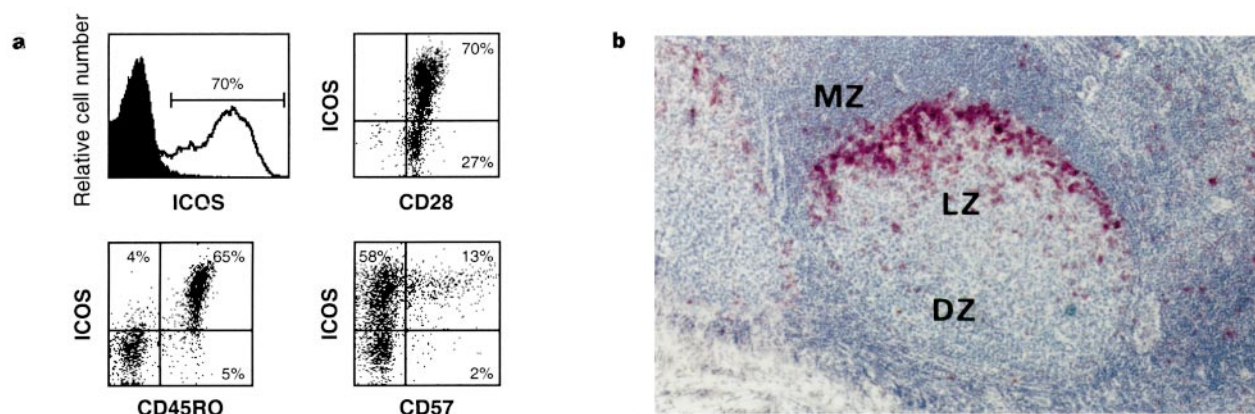


Figure 3 Expression of ICOS on tonsillar T cells. **a**, Tonsillar T lymphocytes were analysed by three-colour flow cytometry on a FACSCalibur (Becton Dickinson). T cells (5×10^4) were gated with a CD3-specific monoclonal antibody (UCHT1-CyChrome, Pharmingen) and expression of ICOS (monoclonal antibody F44-phycoerythrin) was correlated with the expression of CD28 (CD28.2-FITC), CD45RO (UCHL1-FITC), or CD57 (NC1-FITC; all monoclonal antibodies from

Immunotech). Data shown are representative of material obtained from 20 tonsils. **b**, Localization of ICOS $^{+}$ cells in the apical light zone of germinal centres, as determined from histochemical stains of frozen human tonsillar sections with monoclonal antibody F44 using the APAAP technique²⁰. LZ, light zone; DZ, dark zone; MZ, mantle zone (original magnification $\times 69$).

IL-2, whereas ICOS is more effective in superinducing IL-10. It is likely that the interaction of CD28 with its counter-receptors B7-1 and B7-2 is important in the early phases of cooperation between T cells and B cells. At later time points, CTLA-4 counteracts the signals provided by CD28 (refs 6, 9, 18, 19) and thus terminates the production of IL-2; IL-2 is known to drive the expansion of the germinal-centre B cells^{17,20,21}. ICOS, probably binding to receptor(s) on B cells other than B7-1 and B7-2 (see above), may become the dominant signal for T cells in the later phases of T-cell–B-cell cooperation in the lymphoid system. Such a scenario is probable, because IL-10 can induce the terminal differentiation of B cells into memory cells and antibody-secreting plasma cells^{22–24}; this takes place in the apical light zone of the germinal centre^{16,17}, the major site of ICOS expression. □

Methods

Cell preparation, T-cell activation and generation of monoclonal antibodies. Peripheral-blood CD4⁺ (96% pure) and CD8⁺ (92% pure) cells were negatively purified from buffy coats using nylon wool adherence and magnetobead-coupled (Dynal) monoclonal antibodies specific for CD19, CD11b, major histocompatibility complex (MHC) II molecules, and CD8 or CD4. Peripheral-blood CD4⁺ or CD8⁺ T cells (1×10^5 in 200 μ l) were activated using the anti-CD3 monoclonal antibody OKT 3 (ATCC; ascites diluted as indicated in Fig. 2a), and were co-stimulated with monoclonal antibody 9.3 (anti-CD28; 1:3,000 dilution of ascites), F44 (anti-ICOS; 4 μ g ml⁻¹), or MOPC-21 (IgG1 isotype control; 4 μ g ml⁻¹) all of these antibodies were solid-phase-bound through precoating of the 96-well round-bottomed microtitre plates with rabbit anti-mouse serum (10 μ g ml⁻¹, Sigma).

To assay cooperation between T and B cells, peripheral-blood CD4⁺ T cells (50,000 per well) treated with mitomycin C were co-cultured with tonsillar B cells (25,000 per well) in 96-well round-bottomed microtitre plates. Tonsils, obtained from individuals (aged 3–18 yr) undergoing tonsillectomy, were mechanically dispersed and T cells (80–90% pure) were isolated using Ficoll–Hypaque gradient centrifugation followed by nylon-wool adherence. Tonsillar B cells were isolated by rosetting tonsillar cells with sheep red blood cells that had been treated with 2-aminoethyl isothiouraea.

Monoclonal antibodies were generated by fusing spleen cells of BALB/c mice that had been immunized with activated human T cells to myeloma P3X63Ag8.653 (ATCC). The hybridoma secreting the ICOS-specific monoclonal antibody F44 (IgG1) was obtained by subcloning of hybridoma 8F4.

Protein purification and sequencing. MOLT-4V cells (a variant of MOLT-4; ATCC) were lysed in 50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM EDTA, 1 mM phenylmethylsulphonyl fluoride (PMSF) and 1% Nonidet P-40 (NP-40) at 4 °C for 1 h (1 ml buffer per 20×10^6 cells). Nuclei and insoluble material were removed by centrifugation at 1,000g and 100,000g, respectively. The lysate was preabsorbed with unspecific monoclonal antibody coupled to CnBr-activated Sepharose (Pharmacia) and incubated at 4 °C for 4 h with monoclonal antibody F44 coupled to protein G–Sepharose (Pharmacia) according to ref. 25. The F44 matrix was washed in a column with several volumes of buffer 1 (50 mM Tris-HCl, pH 8.0, 300 mM NaCl, 1 mM EDTA, 1 mM PMSF and 0.5% NP-40), buffer 2 (50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1 mM EDTA, 1 mM PMSF, 0.5% NP-40 and 0.1% SDS), and buffer 3 (0.2 M glycine, pH 4.0, and 0.5% CHAPS), and the bound protein was eluted with two volumes of 0.2 M glycine, pH 2.5, and 0.5% CHAPS. The eluate was concentrated by ultrafiltration (Centricon 10, Amicon) and subjected to preparative two-dimensional non-reducing/reducing gel electrophoresis (20×10^9 cell equivalents per gel). After electrophoresis in the first dimension, the gel area containing the ICOS protein was cut out and subjected to reducing conditions in 5.3 M urea, 0.5 M Tris-HCl, pH 8.0, 1% SDS, 1% β -mercaptoethanol (50 °C for 1 h). Free cysteines were subsequently alkylated by addition of 10 mM iodoacetamide (37 °C for 30 min). After equilibration in 1 $\times$ SDS–PAGE sample buffer (30 min), the gel piece was mounted onto a SDS–PAGE slab gel (12% polyacrylamide gel). Following electrophoresis, Coomassie-blue-stained spots corresponding to the ICOS protein (Fig. 2b) were cut out. Protein material from five gels was pooled, *in situ*-digested with trypsin²⁶, fractionated by reverse-phase HPLC on a microRPC C2/C18 SC column (Smart system, Pharmacia) using a linear gradient of acetonitrile and 0.1% trifluoroacetic

acid, and sequenced on a pulsed-liquid gas-phase sequencer (Applied Biosystems).

cDNA cloning. Poly(A)⁺ MOLT-4V RNA was converted to cDNA using oligo-dT primers and Superscript II reverse transcriptase (Gibco), and the cDNA was cloned into λ ZAPII vector (Stratagene) according to the manufacturer's instructions. Peptide XRLTDVT, obtained from protein microsequencing, was used to derive the two degenerate oligonucleotides MGNCTACNGAYGTNAC (512 permutations) and MGNCTDACCNGAYGTNAC (1,024 permutations) for screening of the generated cDNA library by hybridization in 3 M tetramethyl ammonium chloride²⁷. Several positive clones were sequenced using the BigDye Terminator Cycle Sequencing Kit (Applied Biosystems). One of the cDNA clones encoded the above peptide sequence (XRLTDVT) and was used to isolate a full-length cDNA clone.

Received 11 September; accepted 2 December 1998.

- Hara, T., Fu, S. M. & Hansen, J. A. Human T cell activation. II. A new activation pathway used by a major T cell population via a disulfide-bonded dimer of a 44 kilodalton polypeptide (9.3 antigen). *J. Exp. Med.* **161**, 1513–1524 (1985).
- Shahinian, A. et al. Differential T cell costimulatory requirements in CD28-deficient mice. *Science* **261**, 609–612 (1993).
- Lucas, P. J., Negishi, I., Nakayama, K., Fields, L. E. & Loh, D. Y. Naive CD28-deficient T cells can initiate but not sustain an *in vitro* antigen-specific immune response. *J. Immunol.* **154**, 5757–5768 (1995).
- Lenschow, D. J., Walunas, T. L. & Bluestone, J. A. CD28/B7 system of T cell costimulation. *Annu. Rev. Immunol.* **14**, 233–258 (1996).
- Chambers, C. A. & Allison, J. P. Co-stimulation in T cell responses. *Curr. Opin. Immunol.* **9**, 396–404 (1997).
- Walunas, T. L. et al. CTLA-4 can function as a negative regulator of T cell activation. *Immunity* **1**, 405–413 (1994).
- Tivol, E. et al. Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4. *Immunity* **3**, 541–547 (1995).
- Waterhouse, P. et al. Lymphoproliferative disorders with early lethality in mice deficient in Ctla-4. *Science* **270**, 985–988 (1995).
- Thompson, C. B. & Allison, J. P. The emerging role of CTLA-4 as an immune attenuator. *Immunity* **7**, 445–450 (1997).
- Brunet, J. F. et al. A new member of the immunoglobulin superfamily—CTLA-4. *Nature* **328**, 267–270 (1987).
- Peach, R. J. et al. Complementarity determining region 1 (CDR1)- and CDR3-analogous regions in CTLA-4 and CD28 determine the binding to B7-1. *J. Exp. Med.* **180**, 2049–2058 (1994).
- Truneh, A. et al. Differential recognition by CD28 of its cognate counter receptors CD80 (B7.1) and B70 (B7.2): analysis by site directed mutagenesis. *Mol. Immunol.* **33**, 321–334 (1996).
- Nunes, J. et al. CD28 mAbs with distinct binding properties differ in their ability to induce T cell activation: analysis of early and late activation events. *Int. Immunol.* **5**, 311–315 (1993).
- Thompson, C. B. et al. CD28 activation pathway regulates the production of multiple T-cell-derived lymphokines/cytokines. *Proc. Natl Acad. Sci. USA* **86**, 1333–1337 (1989).
- Krocze, R. A. et al. Defective expression of CD40 ligand on T cells causes "X-linked immunodeficiency with hyper-IgM (HIGM1)". *Immunol. Rev.* **138**, 39–59 (1994).
- MacLennan, I. C. Germinal centers. *Annu. Rev. Immunol.* **12**, 117–139 (1994).
- Tsiaghe, V. K., Inghirami, G. & Thorbecke, G. J. The physiology of germinal centers. *Crit. Rev. Immunol.* **16**, 381–421 (1996).
- Krummel, M. F. & Allison, J. P. CTLA-4 engagement inhibits IL-2 accumulation and cell cycle progression upon activation of resting T cells. *J. Exp. Med.* **183**, 2533–2540 (1996).
- Walunas, T. L., Bakker, C. Y. & Bluestone, J. A. CTLA-4 ligation blocks CD28-dependent T cell activation. *J. Exp. Med.* **183**, 2541–2550 (1996).
- Nakanishi, K. et al. Both interleukin 2 and a second T cell-derived factor in EL-4 supernatant have activity as differentiation factors in IgM synthesis. *J. Exp. Med.* **160**, 1605–1621 (1984).
- Jung, L. K., Hara, T. & Fu, S. M. Detection and functional studies of p60-65 (Tac antigen) on activated human B cells. *J. Exp. Med.* **160**, 1597–1602 (1984).
- Rousset, F. et al. Interleukin 10 is a potent growth and differentiation factor for activated human B lymphocytes. *Proc. Natl Acad. Sci. USA* **89**, 1890–1893 (1992).
- Kindler, V. & Zubler, R. H. Memory, but not naive, peripheral blood B lymphocytes differentiate into Ig-secreting cells after CD40 ligation and costimulation with IL-4 and the differentiation factors IL-2, IL-10, and IL-3. *J. Immunol.* **159**, 2085–2090 (1997).
- Choe, J. & Choi, Y. S. IL-10 interrupts memory B cell expansion in the germinal center by inducing differentiation into plasma cells. *Eur. J. Immunol.* **28**, 508–515 (1998).
- Schneider, C., Newman, R. A., Sutherland, D. R., Asser, U. & Greaves, M. F. A one-step purification of membrane proteins using a high efficiency immunomatrix. *J. Biol. Chem.* **257**, 10766–10769 (1982).
- Rosenfeld, J., Capdevielle, J., Guillemot, J. C. & Ferrara, P. In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Anal. Biochem.* **203**, 173–179 (1992).
- Jacobs, K. A. et al. The thermal stability of oligonucleotide duplexes is sequence independent in tetraalkylammonium salt solutions: application to identifying recombinant DNA clones. *Nucleic Acids Res.* **16**, 4637–4650 (1988).
- Aruffo, A. & Seed, B. Molecular cloning of a CD28 cDNA by a high-efficiency COS cell expression system. *Proc. Natl Acad. Sci. USA* **84**, 8573–8577 (1987).
- Williams, A. F. & Barclay, A. N. The immunoglobulin superfamily—domains for cell surface recognition. *Annu. Rev. Immunol.* **6**, 381–405 (1988).
- Cordell, J. L. et al. Immunoenzymatic labeling of monoclonal antibodies using immune complexes of alkaline phosphatase and monoclonal anti-alkaline phosphatase (APAAP complexes). *J. Histochem. Cytochem.* **32**, 219–229 (1984).
- Prediction of transmembrane helices in proteins [online] (cited 01 Dec. 98) (www.cbs.dtu.dk/services/TMHMM-1.0/) (1998).

Acknowledgements. We thank J. Ledbetter for the gift of monoclonal antibody 9.3; J. Slupsky for critical reading of the manuscript; and K. Ranke for technical assistance.

Correspondence and requests for material should be addressed to R.A.K. (e-mail: krocze@rki.de). The ICOS amino-acid sequence has been deposited at the PIR database under accession number S78540.

Direct interaction of microtubule- and actin-based transport motors

Jian-Dong Huang*, Scott T. Brady†, Bruce W. Richards†, David Stenoien†, James H. Resau*, Neal G. Copeland* & Nancy A. Jenkins*

* ABL-Basic Research Program, NCI-Frederick Cancer Research and Development Center, Frederick, Maryland 21702, USA

† Department of Cell Biology and Neuroscience, University of Texas Southwestern Medical Center, Dallas, Texas 75235, USA

The microtubule network is thought to be used for long-range transport of cellular components in animal cells whereas the actin network is proposed to be used for short-range transport¹, although the mechanism(s) by which this transport is coordinated is poorly understood. For example, in sea urchins long-range Ca^{2+} -regulated transport of exocytotic vesicles requires a microtubule-based motor, whereas an actin-based motor is used for short-range transport². In neurons, microtubule-based kinesin motor proteins are used for long-range vesicular transport³ but microtubules do not extend into the neuronal termini, where actin filaments form the cytoskeletal framework⁴, and kinesins are rapidly degraded upon their arrival in neuronal termini⁵, indicating that vesicles may have to be transferred from microtubules to actin tracks to reach their final destination. Here we show that an actin-based vesicle-transport motor, MyoVA (ref. 6), can interact directly with a microtubule-based transport motor, KhcU. As would be expected if these complexes were functional, they also contain kinesin light chains and the localization of MyoVA and

KhcU overlaps in the cell. These results indicate that cellular transport is, in part, coordinated through the direct interaction of different motor molecules.

In rats and mice, MyoVA is encoded by the *dilute* locus^{7,8}. *dilute*-null animals show defects in melanosome transport^{9,10} and organization of smooth endoplasmic reticulum (SER) in cerebellar Purkinje cells^{11,12}. To identify cellular proteins that interact directly with MyoVA, and which may, therefore, help to facilitate or coordinate MyoVA-associated vesicle transport, we used a yeast two-hybrid screen to look for MyoVA-interacting proteins. We used as baits three protein fragments (numbers 10–12) that span the AF-6/cno-homology region of the MyoVA tail (Fig. 1a). We chose this region because it may be important in vesicle transport and may serve as a protein–protein interaction site^{13–15}. Two of the baits (10 and 12) screened against a library of mouse brain complementary DNAs fused to the *GAL4* transcriptional activation domain or against a library of mouse embryo cDNAs fused to VP16 did not give rise to any positive clones. In contrast, bait 11 (amino acids 1,569–1,768), which covers most of the AF-6/cno-homology domain (amino acids 1,527–1,805), gave rise to 14 doubly positive clones ($\text{His}^+/\beta\text{-Gal}^+$), 7 of which contained a portion of the mouse ubiquitous kinesin heavy chain (KhcU)¹⁶. Although this result was unexpected, because KhcU is a microtubule-based motor that has been implicated in the transport of membranous organelles^{17–21}, a direct interaction between MyoVA and KhcU could provide a basis for coordinating actin- and microtubule-based organelle transport.

All seven KhcU-positive clones have unique ends but overlap with each other (Fig. 1b). The common region of overlap (S763–D856) is located at the end of the rod domain, near the junction with the tail. One likely function of the rod domain is to act as a lever arm for motility, whereas the more distal regions are thought to interact with the kinesin light chains and with membrane surfaces. Our results indicate that another function of the distal rod domain may be to interact with a myosin motor. The MyoVA-binding site

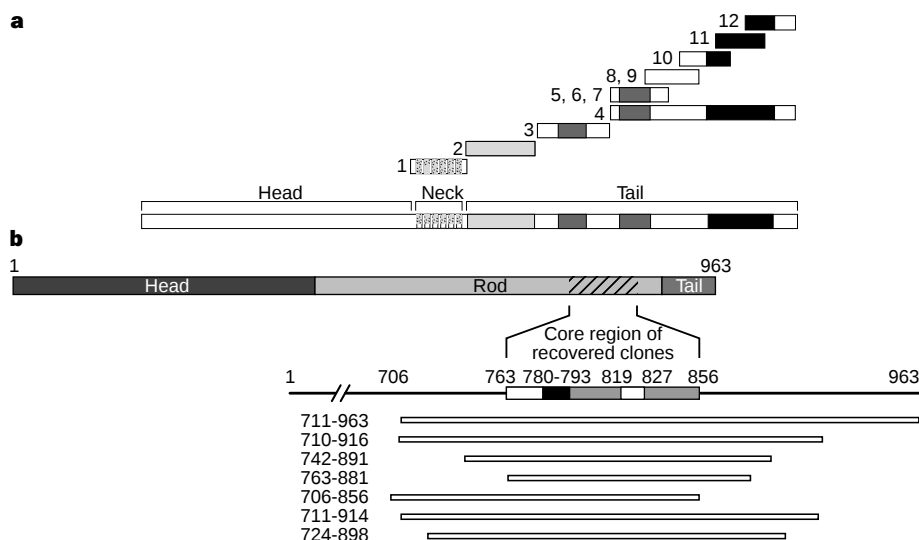


Figure 1 Domains of MyoVA, showing the baits (numbered 1–12) used in yeast two-hybrid screens and the region of MyoVA that interacts with KhcU. **a**, Bottom, the different functional domains of MyoVA, including the head, neck and tail. The six light-chain-binding IQ-type motifs in the neck are shown as stippled boxes, and three coiled-coil dimerization domains and the AF-6/cno-homology domain are shown as grey and black rectangles, respectively. Top, the fragments of MyoA that were used as baits in yeast two-hybrid screens. Baits 5–9 contain different alternative splice forms of the MyoVA tail that are differentially expressed in brain and in skin (melanocytes)³⁰. **b**, Top, the different functional domains of KhcU, including the head, rod and tail. Bottom, the relative positions of the KhcU clones identified in the yeast two-hybrid screen. Middle, the core region (amino acids

763–856), which is overlapped by all seven clones. The core region is composed of non-conserved sequences (open rectangles), sequences that are conserved among conventional kinesins (grey rectangles), and sequences that are highly conserved and found in kinesins even as distantly related to KhcU as is bimC (black rectangle). Three domains can be defined within the core overlap region: two stretches contain heptad repeats that are characteristic of coiled-coil structure (amino acids 763–792 and 829–856) and are interrupted by a third stretch (amino acids 793–828) that does not contain heptads. The first of the two non-conserved sequences retains the heptad pattern (amino acids 763–780), but the second such sequence (amino acids 819–827) is part of the interrupting gap.

is conserved among conventional kinesins in vertebrates and *Drosophila*²² and a 14-amino-acid stretch (amino acids 780–793) within the core of the binding site is also found in a kinesin-related protein, bimC, from *Aspergillus*²³. The common region of overlap in the MyoVA-binding site of KhcU centres on a break in the heptad-repeat motifs of the stalk (amino acids 793–828). On the carboxy-terminal side of this break, the heptads are thought to form a coiled-coil α -helix consisting of a light-chain/heavy-chain heterodimer, whereas the amino-terminal side of the break represents portions of the stalk formed by kinesin-heavy-chain homodimeric coiled-coil α -helix (S.T.B. and B.W.R., unpublished observations). As a result, the globular N terminus of the kinesin light chain (amino acids 1–85) and this gap in the kinesin-heavy-chain heptads (amino acids 793–828) would be closely apposed to each other.

To determine whether other regions of the MyoVA tail bind KhcU, we used a portion of the KhcU rod–tail domain (amino acids 711–963) in a yeast two-hybrid assay with a test panel of unrelated proteins and the 12 MyoVA baits described in Fig. 1a. Only bait 11 supported growth on selective media (Fig. 2a), although all clones grew well on non-selective media (data not shown). Surprisingly, MyoVA bait 4, which contains all of bait 11 (Fig. 1a), failed to support growth on selective media (Fig. 2a). The fusion protein of relative molecular mass 80,000 (M_r 80K) encoded by bait 4 may have a conformation in yeast that precludes its interaction with KhcU, or perhaps it contains sequences that prevent entry into the nucleus.

In a yeast two-hybrid assay, we identified only one kinesin among the large family of kinesins identified so far that was capable of interacting with MyoVA, indicating that MyoVA might interact specifically with KhcU. In agreement with this idea, neuronal kinesin heavy chain (KhcN), a protein closely related to KhcU, cannot interact functionally with MyoVA in a yeast two-hybrid assay

(Fig. 2b). The difference in binding between KhcU and KhcN indicates that one or both of the two small non-conserved regions in the MyoVA-binding site of KhcU is important for binding specificity. The first of these (amino acids 763–774) is part of the heptads in the stalk that are important for heavy-chain/heavy-chain interactions. Such domains may also be involved in protein–protein interactions, for example in the formation of thick filaments from myosin II in muscle. Even more interesting, however, is the fact that the second non-conserved region extends from amino acids 819 to 827; this region is part of the central gap in KhcU heptads and should be near the kinesin-light-chain N terminus. Thus, for the KhcU domain that binds MyoVA, the amino acid 819 to 827 stretch is the most divergent from the equivalent region of KhcN and represents a strong candidate for specifying the binding of MyoVA to KhcU, but not to KhcN.

MyoVA and KhcU can also bind to each other *in vitro*. We incubated similar amounts of five different fusion proteins consisting of glutathione-S-transferase (GST) and MyoVA tail regions (Fig. 2c, lower panel; bait 12 did not express stable protein) with KhcU-coated resin and washed the proteins; we then analysed the bound proteins by SDS–polyacrylamide gel electrophoresis (PAGE) and immunoblotted the gels with anti-GST antibody. The two MyoVA-fusion proteins containing the AF-6/cno-homology domain (baits 4 and 11) bound to the resin (Fig. 2c, upper panel) but, as expected, baits that did not contain the AF-6/cno-homology domain (baits 5, 8 and 10) did not.

MyoVA and KhcU also interact in a mammalian two-hybrid screen, a very stringent test for authentic interaction partners²⁴. In these experiments, the MyoVA tail (bait 11) was fused to the DNA-binding domain of Gal4 and the KhcU rod–tail domain was fused to the activator domain of VP16. When these two constructs

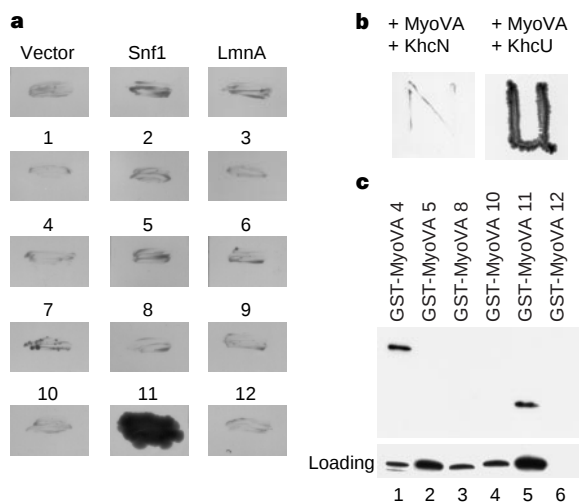


Figure 2 Specificity of the interaction between MyoVA and KhcU. **a**, Yeast strains expressing the KhcU rod–tail domain and one of the twelve Gal4–MyoVA fusion proteins (containing the MyoVA domains shown in Fig. 1a), vector alone, Gal4–Snf1 (yeast serine threonine protein kinase 1), or Gal4–LmnA (human lamin A/C) were grown on selective media (AA-Trp-Leu-His + 25 mM 3-AT) to test the specificity of the interaction between KhcU and MyoVA. Only bait 11, which contains the AF-6/cno-homology domain, supported growth on selective media. **b**, Plasmids containing the KhcU rod–tail domain or the KhcN (GenBank accession number X61435) rod–tail domain, fused to VP16, were introduced into yeast strains expressing the MyoVA bait 11 tail fused to the Gal4 activation domain. The yeast patch containing the KhcN rod–tail did not grow on selective media. **c**, Two GST–MyoVA fusion fragments, expressed in *E. coli* (lower panel), containing the AF-6/cno-homology domain (baits 4 and 11) bound to S*Bind resin coated with S-tagged KhcU rod–tail protein (upper panel), whereas three fragments lacking the homology domain (baits 5, 8 and 10) failed to bind the resin.

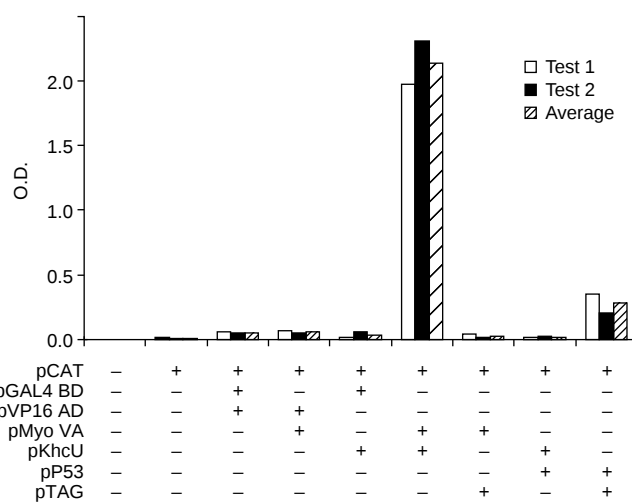


Figure 3 MyoVA interacts with KhcU in mammalian cells. We assayed the ability of MyoVA to interact with KhcU in mammalian cells by co-transfecting expression constructs containing the MyoVA tail (bait 11) fused to a Gal4 DNA-binding domain (pMyoVA), the KhcU rod–tail fused to the VP16 activator domain (pKhcU), and a CAT reporter (pCAT) into mammalian cells and then assaying the level of CAT activity 65 h later. T antigen fused to the Gal4 DNA-binding domain (pTAG) and p53 fused to the VP16 activator domain (pP53) were co-transfected as a positive control. Significant levels of CAT activity were observed only in the positive control or in cells receiving both pMyoVA and pKhcU. pGAL4 BD and pVP16 AD are plasmids containing only the binding domain of Gal4 or the activator domain of VP16, respectively.

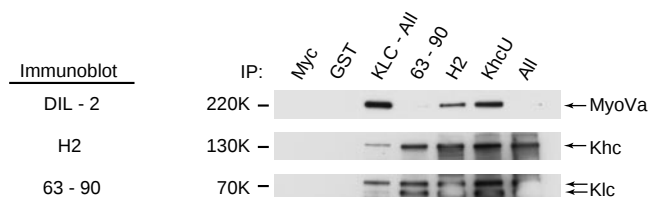


Figure 4 MyoVA co-immunoprecipitates with KhcU and Klc (kinesin light chain) in normal cell extracts. Western blots of brain extracts precipitated with four monoclonal anti-kinesin antibodies (KLC-All, 63-90, H2 and KhcU; see text), either alone or in combination, or with anti-Myc or anti-GST antibodies as controls, were probed with antibodies specific for MyoVA (DIL-2; top), KhcU (H2; middle), or Klc (63-90; bottom). IP, immunoprecipitate.

were co-transfected into mammalian cells, under the control of the SV40 early promoter, along with a chloramphenicol acetyltransferase (CAT) reporter gene, strong CAT activity was observed (Fig. 3). The level of CAT activity was, in fact, much greater than in the positive control cells (which were transfected with p53 and T antigen). In contrast, neither construct generated any significant CAT activity when transfected alone.

Next, we asked whether MyoVA would co-immunoprecipitate with KhcU from normal mouse brain extracts, and whether the complexes also contained kinesin light chains, as would be expected if they were functional motors. We used four different monoclonal antibodies for the co-immunoprecipitation experiments, namely H2 and KhcU for kinesin heavy chains, and 63-90 and Klc-All for kinesin light chains. H2 reacts with the motor-neck region of all kinesin heavy chains (KhcU and KhcN)²⁰; KhcU is specific for KhcU and recognizes an epitope at the C terminus of KhcU (D.S., S.T.B. and B.W.R., unpublished observations); 63-90 recognizes kinesin light chains 1 and 2 at a conserved epitope near the N terminus; and Klc-All recognizes all kinesin light chains at a conserved epitope in the tandem repeat domain²⁵. When brain extracts were immunoprecipitated with H2, KhcU or Klc-All, the immunoprecipitates contained MyoVA immunoreactivity as well as kinesin heavy and light chains (Fig. 4). As expected, no MyoVA immunoreactivity was detected when extracts were made from mice carrying a null mutation in *Myo5a* (data not shown). Immunoprecipitates obtained with the 63-90 antibody contained very little MyoVA immunoreactivity, as did immunoprecipitates made using a combination of all four antibodies. These results are interesting, because the 63-90 epitope maps to a conserved region (amino acids 1–50) of the kinesin light chain²⁵. As indicated above, this light-chain domain is next to the region of KhcU (amino acids 763–856) that binds to the MyoVA tail.

Finally, the results of immunofluorescence experiments using a

mouse melanocyte cell line (Melan-a) and antibodies specific for MyoVA and KhcU indicated that these two proteins often co-localize in the cell. Co-localization was particularly evident in the perinuclear region and the dendritic arbor (Fig. 5). In the cytoplasm, both MyoVA and KhcU form a punctate staining pattern that partially, but not completely, overlaps. The results that we obtained with anti-MyoVA antibodies are similar to previously published results^{9,10} but differ in that we used a largely amelanotic melanocyte cell line; thus our cells do not give the highly punctate staining pattern that is indicative of melanosome staining. Our results are also consistent with previous studies that showed that MyoVA-associated organelles are located on both microtubule and actin filaments in superior cervical ganglion growth cones²⁶ and with our unpublished studies showing that transient overexpression of His-tagged KhcU rod-tail protein in COS-7 cells is sufficient to recruit more MyoVA onto microtubules.

Considerable amounts of cell biological and biochemical evidence indicate that long-range transport in animal cells may occur on microtubules whereas short-range transport occurs on actin filaments. An important question raised by such studies is how transport along these two cellular highways is coordinated. Our results support the theory that this transport is coordinated in part through the direct interaction of motor molecules from the different transport networks. These data are consistent with previous studies that showed that MyoVA and kinesin can both bind synaptic vesicles in the rat^{17,27} and that MyoVA is required for short-range, but not long-range, transport of SER in Purkinje cells^{11,12}. Yeast studies have showed that Smy1 (a kinesin-related protein) can suppress the phenotype produced by mutation of Myo2 (a yeast homologue of mammalian MyoVA)^{28,29} and that Myo2 and Smy1 co-localize to the same regions of polarized growth in the cell²⁹, further indicating that MyoVA motors may be able to interact with microtubule-based motors.

Although the interaction between MyoVA and KhcU could coordinate transport in several different ways, one appealing model proposes that MyoVA and KhcU bind to the same cargo, as well as to each other, to form an integrated motor complex. In this model, the activity of the complex is tightly regulated such that only one motor in the complex is dominant at any given time. The advantage of such a motor complex is that it would allow a single vesicle to be transported by multiple motors and to move on both microtubule and actin tracks. A vesicle with multiple motors can make the transition from microtubule to actin filaments easily by switching motor activity. In microtubule-rich cellular domains, the greater reach of KhcU would lead to long-range microtubule-based movements. In regions such as the cellular cortex, the neuronal growth cone or the presynaptic terminus, where microtubules are rare or absent, MyoVA would take over to deliver the cargo to a final destination. Such switches in motility are likely to be essential in microtubule-deficient regions such as these. □

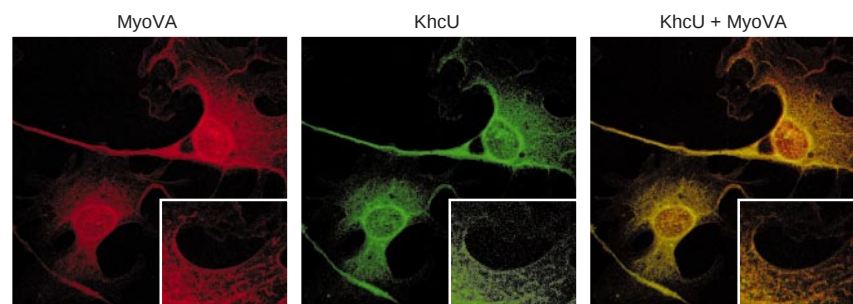


Figure 5 MyoVA and KhcU proteins often co-localize in Melan-a cells. Melan-a cells were stained with anti-MyoVA antibody (DIL-2) or a KhcU-specific monoclonal antibody (KhcU). Left, MyoVA immunofluorescence (red); centre, KhcU

immunofluorescence (green); right, superimposition of the two images. A higher magnification of each image is shown in the insets.

Methods

Plasmid construction. DNA inserts for most plasmid constructs were amplified by the polymerase chain reaction (PCR) using the Expand High Fidelity PCR system (Boehringer) and primers linked to restriction enzyme sites suitable for cloning. Inserts were sequenced to rule out possible mutations introduced by PCR.

The bait plasmids were constructed by insertion of PCR-amplified MyoVA fragments into the *NcoI*–*SalI* sites of the pAS2 vector (a gift from S. J. Elledge). The 5' and 3' primers used for PCR amplification were, respectively: bait 1, 5'-GTGGCCATGGGCATCTTTTCCGTGCTGGTCAAG and 5'-GGGCGTCGACGTTATCTCCATGCAATATGGAGCTTC; 2, 5'-GTGGCCATGGGCCTCAAAATTGAGGCTCGCTCTG and 5'-GGGCGTCGACGTTATTAGGCACATTCAGCATCAGGG; 3, 5'-GTGGCCATGGGCCAGGACACAAGAGAACA-GAC and 5'-GGGCGTCGACGTTATCACTTTGTCTTTCATTTCTG; 4, 5'-GTGGCCATGGGCATCTTGAGGTGCGACGCTGGTG and 5'-GGGCGTCGACGTTATCAGTAGCTCACTGGAACAAC; 5–7, 5'-GTGGCCATGGGCATCTTGAGGTGCGACGCTGGTG and 5'-GGGCGTCGACGTTATTTCGGGGAATGTTGACTGGC; 8, 9, 5'-GTGGCCATGGGCGCCCGCATTGAGGCCAGCCTG and 5'-GGGCGTCGACGTTACTTCAAACAGTGCAAAAATCGAC; 10, 5'-GTGGCCATGGGCGAAAAGGATTTCCAAGGAATG and 5'-GGGCGTCGACGTTAGTTGAGTGCCTGAAGAATGGAGTC; 11, 5'-GTGGCCATGGCCAAATAGTGGAGAGAGGGG and 5'-GTGGCCATGGGCTCCTTC-CACCTCAGTCATGTGTC; 12, 5'-GTGGCCATGGGCTCCTTCCACTCAGTC-ATGTGTC and 5'-GGGCGTCGACGTTATCAGTAGCTCACTGGAACAAC.

For mammalian two-hybrid assays, plasmids containing cDNAs encoding the Gal4 DNA-binding domain (pM), the VP16 activation domain (pVP16) and the CAT reporter (pG5CAT, called pCAT here) were purchased from Clontech. pMyoVA was constructed by inserting MyoVA fragment 11 into pM. pKhcU was made by cloning a KhcU gene fragment (encoding amino acids 711–963; Fig. 1) into pVP16.

Bacterial expression plasmids were made by fusing cDNAs encoding MyoVA fragments from the pAS2 plasmids into the GST expression vector pGEX 4T-1 (Pharmacia). The KhcU cDNA fragment described above was inserted into pET 32 (+) (Novagene).

Cell lines and reagents. Rabbit anti-GST antibody was purchased from Boehringer. Rabbit anti-MyoVA antibody, DIL-2, was provided by J. Hammer¹⁰. Fluorescein isothiocyanate (FITC)-conjugated AffiniPure goat anti-mouse IgG (H+L) was from Jackson ImmunoResearch Laboratories. Tetramethylrhodamine goat anti-rabbit IgG (H+L) conjugate was from Molecular Probes. **Yeast two-hybrid screen and β -galactosidase assay.** The mouse brain cDNA library was from Clontech; the mouse embryo cDNA library, consisting of cDNAs from embryos at 9.5 and 10.5 days post-coitum, was constructed by S. Hollenberg. Library screen and β -galactosidase assays were performed according to Clontech protocol.

Mammalian two-hybrid assay. The mammalian two-hybrid assay was done according to Clontech protocol. CAT activity was measured with the CAT enzyme-linked immunosorbent-assay kit (Boehringer).

In vitro binding and co-precipitation analysis. Induction of *Escherichia coli* strain BL21(DE3) containing various expression plasmids by isopropyl- β -D-thiogalactoside resulted in overexpression of the fusion proteins encoded by the expression plasmids. Protein extracts were made from the induced cultures by sonication as described in product manuals (Novagen and Pharmacia). GST–MyoVA fusion proteins were purified with glutathione–Sephacrose 4B beads (Pharmacia Biotech) according to the manufacturer's protocol (GST gene fusion system, Pharmacia Biotech). S*Bind resins were coated by incubating them with KhcU extracts at room temperature for 30 min; the resins were then washed three times with 1 $\times$ PBS. GST-fusion proteins were incubated with S*KhcU-coated S*Bind resins for 2 h at 4°C in 1 $\times$ PBS + 0.1% Tween 20 and washed four times with 1 $\times$ PBS + 0.1% Tween 20 at 4°C for 15 min each time. Retained proteins were analysed by SDS–PAGE and detected by autoradiography or with a GST western blotting kit (Boehringer).

Immunoblots and immunoprecipitation. The H2 (ref. 20), Klc–All and 63–90 (ref. 25) antibodies have been characterized previously. The KhcU antibody was raised against a synthetic peptide in the KhcU tail that was not present in KhcN, injected into mice and used to generate monoclonal antibodies as described²⁰.

Immunoprecipitations and immunoblots were done according to published methods²⁵.

Immunofluorescence. Cells plated on Chamber Slides (Nalge Nunc International) were fixed after 16–24 h for 10 min with methanol (–20°C), washed with 1 $\times$ PBS, and blocked for 30 min in two changes of 1 $\times$ PBS containing 10% normal goat serum. Cells were then incubated for 1 h at room temperature with primary antibodies diluted in PBS containing 0.2% saponin (Sigma). The cells were then washed briefly with PBS, blocked again for 30 min and incubated for 1 h at room temperature with secondary antibodies. After a brief wash with PBS, the slides were mounted with glass coverslips using GEL/MOUNT (Biomed).

Received 17 September; accepted 23 November 1998.

- Langford, G. M. Actin- and microtubule-dependent organelle motors: interrelationships between the two motility systems. *Curr. Opin. Cell Biol.* **7**, 82–88 (1995).
- Bi, G. Q. *et al.* Kinesin- and myosin-driven steps of vesicle recruitment for Ca^{2+} -regulated exocytosis. *J. Cell Biol.* **138**, 999–1008 (1997).
- Brady, S. T. A kinesin medley: biochemical and functional heterogeneity. *Trends Cell Biol.* **5**, 159–164 (1995).
- LeBeux, Y. J. & Willemot, J. An ultrastructural study of the microfilaments in rat brain by means of E-PTA staining and heavy meromyosin labeling. II. The synapses. *Cell Tissue Res.* **160**, 37–68 (1975).
- Okada, Y., Yamazaki, H., Sekine-Aizawa, Y. & Hirokawa, N. The neuron-specific kinesin superfamily protein KIF1A is a unique monomeric motor for anterograde axonal transport of synaptic vesicle precursors. *Cell* **81**, 769–780 (1995).
- Cheney, R. E. *et al.* Brain myosin-V is a two-headed unconventional myosin with motor activity. *Cell* **75**, 13–23 (1993).
- Dekker-Ohno, K., Oda, S., Yamamura, H. & Kondo, K. An ataxic mutant rat with dilute coat color. *Lab. Anim.* **33**, 370–372 (1993).
- Mercer, J. A., Seperack, P. K., Strobel, M. C., Copeland, N. G. & Jenkins, N. A. A novel myosin heavy chain encoded by murine dilute coat colour locus. *Nature* **349**, 709–713 (1991).
- Provance, D. W., Wei, M., Ipe, V. & Mercer, J. A. Cultured melanocytes from dilute mutant mice exhibit dendritic morphology and altered melanosome distribution. *Proc. Natl Acad. Sci. USA* **93**, 14554–14558 (1996).
- Wu, X. F., Bowers, B., Wei, Q., Kocher, B. & Hammer, J. A. Myosin V associates with melanosomes in mouse melanocytes—evidence that myosin V is an organelle motor. *J. Cell Sci.* **110**, 847–859 (1997).
- Dekker-Ohno, K. *et al.* Endoplasmic reticulum is missing in dendritic spines of Purkinje cells of the ataxic mutant rat. *Brain Res.* **714**, 226–230 (1996).
- Takagishi, Y. *et al.* The dilute-lethal (*dl*) gene attacks a Ca^{2+} store in the dendritic spine of Purkinje cells in mice. *Neurosci. Lett.* **215**, 169–172 (1996).
- Prasad, R. *et al.* Cloning of the ALL-1 fusion partner, the AF-6 gene, involved in acute myeloid leukemias with the t(6;11) chromosome translocation. *Cancer Res.* **53**, 5624–5628 (1993).
- Miyamoto, H. *et al.* canoe encodes a novel protein containing a GLGF/DHR motif and functions with Notch and scabrous in common developmental pathways in *Drosophila*. *Genes Dev.* **9**, 612–625 (1995).
- Ponting, C. P. AF-6/cno: neither a kinesin nor a myosin, but a bit of both. *Trends Biochem. Sci.* **20**, 265–266 (1995).
- Gudkov, A. V. *et al.* Cloning mammalian genes by expression selection of genetic suppressor elements: association of kinesin with drug resistance and cell immortalization. *Proc. Natl Acad. Sci. USA* **91**, 3744–3748 (1994).
- Leopold, P. L., McDowell, A. W., Pfister, K. K., Bloom, G. S. & Brady, S. T. Association of kinesin with characterized membrane-bounded organelles. *Cell Motil. Cytoskeleton* **23**, 19–33 (1992).
- Brady, S. T., Pfister, K. K. & Bloom, G. S. A monoclonal antibody against kinesin inhibits both anterograde and retrograde fast axonal transport in squid axoplasm. *Proc. Natl Acad. Sci. USA* **87**, 1061–1065 (1990).
- Hirokawa, N. *et al.* Kinesin associates with anterogradely transported membranous organelles *in vivo*. *J. Cell Biol.* **114**, 295–302 (1991).
- Pfister, K. K., Wagner, M. C., Stenoien, D. L., Brady, S. T. & Bloom, G. S. Monoclonal antibodies to kinesin heavy and light chains stain vesicle-like structures, but not microtubules, in cultured cells. *J. Cell Biol.* **108**, 1453–1463 (1989).
- Rodionov, V. I., Gyoova, E. K. & Gelfand, V. I. Kinesin is responsible for centrifugal movement of pigment granules in melanophores. *Proc. Natl Acad. Sci. USA* **88**, 4956–4960 (1991).
- Niclas, J., Navone, F., Hom-Booher, N. & Vale, R. D. Cloning and localization of a conventional kinesin motor expressed exclusively in neurons. *Neuron* **12**, 1059–1072 (1994).
- Navone, F. *et al.* Cloning and expression of a human kinesin heavy chain gene: interaction of the COOH-terminal domain with cytoplasmic microtubules in transfected CV-1 cells. *J. Cell Biol.* **117**, 1263–1275 (1992).
- Tsan, J. T. *et al.* In *The Yeast Two-Hybrid Systems* (eds Bartel, P. L. & Fields, P. S.) 217–232 (Oxford Univ. Press, Oxford, 1997).
- Stenoien, D. S. & Brady, S. T. Immunochemical analysis of kinesin light chain function. *Mol. Biol. Cell* **8**, 675–689 (1997).
- Evans, L. L., Hammer, J. & Bridgman, P. C. Subcellular localization of myosin V in nerve growth cones and outgrowth from dilute-lethal neurons. *J. Cell Sci.* **110**, 439–449 (1997).
- Prekeris, R. & Terrian, D. M. Brain myosin V is a synaptic vesicle-associated motor protein: evidence for a Ca^{2+} -dependent interaction with the synaptobrevin-synaptophysin complex. *J. Cell Biol.* **137**, 1589–1601 (1997).
- Lillie, S. H. & Brown, S. S. Suppression of a myosin defect by a kinesin-related gene. *Nature* **356**, 358–361 (1992).
- Lillie, S. H. & Brown, S. S. Immunofluorescence localization of the unconventional myosin, Myo2p, and the putative kinesin-related protein, Smylp, to the same regions of polarized growth in *Saccharomyces cerevisiae*. *J. Cell Biol.* **125**, 825–842 (1994).
- Seperack, P. K., Mercer, J. A., Strobel, M. C., Copeland, N. G. & Jenkins, N. A. Retroviral sequences located within an intron of the dilute gene alter dilute expression in a tissue-specific manner. *EMBO J.* **14**, 2326–2332 (1995).

Acknowledgements. We thank J. Hammer for the MyoVA antibody DIL-2; S. J. Elledge for the pAS2 vector and Y187 yeast strain; S. Hollenberg for the mouse embryonic cDNA library; and R. Frederickson for help with graphics. This work was supported by the National Cancer Institute, DHHS, under contract with ABL, and by a grant to S.T.B. from the National Institute of Neurological Disease and Stroke NINDS, a joint grant from NASA and the National Institute of Aging, and a grant from the Welch Foundation.

Correspondence and requests for materials should be addressed to N.A.J. (e-mail: jenkins@ncicfcrf.gov).

Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase

Heather P. Harding, Yuhong Zhang & David Ron

Skirball Institute of Biomolecular Medicine, Departments of Medicine and Cell Biology and the Kaplan Cancer Center, NYU School of Medicine, New York, New York 10016, USA

Protein synthesis and the folding of the newly synthesized proteins into the correct three-dimensional structure are coupled in cellular compartments of the exocytosis pathway by a process that modulates the phosphorylation level of eukaryotic initiation factor-2 α (eIF2 α) in response to a stress signal from the endoplasmic reticulum (ER)^{1,2}. Activation of this process leads to reduced rates of initiation of protein translation during ER stress³. Here we describe the cloning of *perk*, a gene encoding a type I transmembrane ER-resident protein. PERK has a luminal domain that is similar to the ER-stress-sensing luminal domain of the ER-resident kinase Ire1, and a cytoplasmic portion that contains a protein-kinase domain most similar to that of the known eIF2 α kinases, PKR and HRI. ER stress increases PERK's protein-kinase activity and PERK phosphorylates eIF2 α on serine residue 51, inhibiting translation of messenger RNA into protein. These properties implicate PERK in a signalling pathway that attenuates protein translation in response to ER stress.

In eukaryotic cells, the folding of proteins that are destined to be secreted or membrane-bound takes place in the ER. This process is impeded when cells are deprived of essential nutrients or energy sources or exposed to toxins that perturb the specialized environment of the ER⁴. The accumulation of incorrectly folded proteins triggers an ER-stress response that has at least two distinct components. The first, known as the unfolded-protein response (UPR), consists of the transcriptional induction of genes that encode ER-resident proteins such as Grp78, Grp94 and protein disulphide isomerase. These proteins are believed to promote the folding of newly synthesized peptides in the ER lumen^{5,6}. The second component consists of a profound and rapid repression of protein synthesis³. Both responses can be rationalized in terms of the need to relieve ER stress: the first response increases the capacity of the ER to actively fold proteins and the second decreases the demands made on the organelle by attenuating protein-synthesis rates².

The mediators of the UPR have been well characterized, first in yeast^{7,8} and, more recently, in mammalian cells^{9,10}. The sensor of ER stress in both types of cell is an ER-resident transmembrane kinase, IRE1, which transduces a luminal signal imparted by the presence of incorrectly folded proteins to a nuclear event that results in the increased transcription of the previously mentioned genes¹¹. Reduced protein synthesis in response to ER stress is associated with polysome disassembly and correlates with increased phosphorylation of eIF2 α (refs 1, 3). Phosphorylated eIF2 α interferes with the formation of an active 43S translation-initiation complex¹² and expression of a non-phosphorylatable mutant eIF2 α (with a mutation of residue S51 to A; S51A) attenuates translational inhibition by ER stress¹³. Two different eIF2 α kinases have been identified in mammalian cells, haem-regulated eIF2 α kinase (HRI)¹⁴ and the interferon-inducible RNA-dependent protein kinase (PKR)¹⁵. Ca²⁺ release from ER stores enhances the kinase activity of PKR¹⁶ and dominant-negative mutant forms of PKR attenuate the inhibition of translation mediated by Ca²⁺ release¹³. Although both of these results are consistent with a role for PKR in translational inhibition

by ER stress, the level of inhibition of protein synthesis is nearly the same in wild-type and *pkrr*^{-/-} cells (see below). This suggests the existence of alternative eIF2 α kinases that may respond to ER stress.

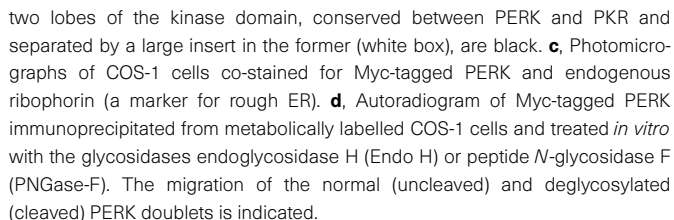
A search of the nucleotide databases (GenBank) identified a *Caenorhabditis elegans* cosmid, CEF46C3, that contains a gene encoding a predicted type I transmembrane protein (one with its amino terminus in the lumen) with a cytoplasmic portion that has a protein-kinase domain most similar to that of PKR and HRI and an external/luminal domain that shares blocks of identity with *C. elegans* and mammalian Ire1. A human expressed sequence tag (EST) that encodes a peptide fragment with similarity to this *C. elegans* protein was used as a hybridization probe to clone a full-length complementary DNA encoding a murine homologue. The murine protein, of 1,114 residues, is ubiquitously expressed and has the same predicted type I transmembrane topology and ~30% overall identity to the *C. elegans* protein. Its N terminus has blocks of identity with mammalian Ire1, totalling 20% identity, and its carboxy terminus is ~40% identical to PKR (Fig. 1a, b). A C-terminal Myc-tagged derivative of the full-length protein, when expressed in COS-1 cells, co-localized with endogenous ribophorin (an ER-membrane marker; Fig. 1c). Increased mobility of the metabolically labelled protein doublet in immunoprecipitates treated with either endoglycosidase H or peptide N-glycosidase F indicates that it is a glycoprotein that is not transported to the distal Golgi complex and resides in the ER. We do not know why there are two species of the protein in the transfected cells (Fig. 1d). On the basis of its similarity to PKR and its localization in the ER membrane, we named the protein PERK (for PKR-like ER kinase). We proposed that PERK, like Ire1, might respond to ER-stress signals and transduce these to changes in protein-synthesis rates through phosphorylation of eIF2 α .

The purified kinase domain of PERK, expressed as a fusion protein with glutathione-S-transferase (GST) in *Escherichia coli* undergoes autophosphorylation *in vitro* (Fig. 2a). The active kinase phosphorylates purified eIF2 α effectively, indicating that the latter is a direct substrate (Fig. 2a, middle). Phosphorylation of eIF2 α takes place on S51 (Fig. 2b), the same residue that is targeted by PKR and HRI¹². Treatment of translation-competent reticulocyte lysate with purified, bacterially expressed PERK leads to a profound inhibition in mRNA translation. Replacing K618 of PERK with alanine (K618A), a mutation that abolishes the ability of the protein to undergo autophosphorylation or to phosphorylate eIF2 α (Fig. 2a, lane 2, top and middle), also abolishes its ability to inhibit translation (Fig. 2c, lane 2). Cells transfected with PERK expression plasmids accumulate large quantities of the protein, and when levels of expression reach a certain threshold PERK's kinase activity is activated (Fig. 3c, compare lanes 2, 9). This correlates with profound inhibition of mRNA translation (Fig. 2d, e), indicating that the effects of PERK on protein translation observed *in vitro* in the reticulocyte lysate also occur in cells *in vivo*.

Mouse embryonic fibroblasts that are nullizygous for *pkrr* are not impaired in their ability to reduce levels of protein synthesis when treated with agents that induce ER stress (Fig. 3a). This indicates that another eIF2 α kinase, perhaps PERK, may be involved in the attenuation of translation under these circumstances. Results of metabolic labelling experiments show that PERK is a phosphoprotein. In response to ER stress, it undergoes hyperphosphorylation and an associated shift in mobility during SDS-polyacrylamide-gel electrophoresis (Fig. 3b). The shift is not observed for the inactive K618A mutant, is reduced by dephosphorylation *in vitro* and correlates with the autokinase activity of immunoprecipitated PERK (Fig. 3b, c). These ER-stress-induced changes in PERK are consistent with intermolecular or intramolecular autophosphorylation, as has been suggested for the activation of the other known eIF2 α kinases, HRI¹⁷ and PKR¹⁵. The stress-activated conformational shift in PERK is independent of synthesis of new PERK protein (Fig. 3c, lane 4), and treatment with arsenite, heat-shock or

The structural similarity between PERK and Ire1 indicates that both proteins may use a similar mechanism to transduce ER stress. This theory is supported by the observation that overexpression of

The UPR and translational inhibition by ER stress proceed by



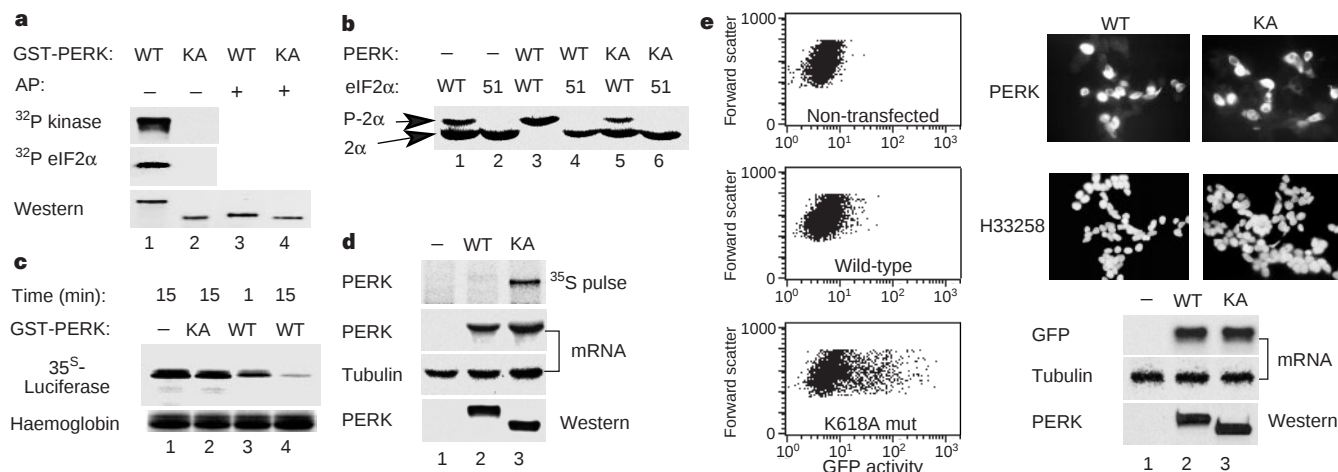


Figure 2 PERK is an eIF2 α kinase that inhibits translation. **a**, Top, autoradiogram of the *in vitro* autokinase activity of the wild-type (WT) and K618A mutant (KA) PERK, purified as GST-fusion proteins from *E. coli*. Middle, autoradiogram of eIF2 α that had been added to the kinase reaction and immunoprecipitated with a specific antiserum. Bottom, PERK western blot of same lysate untreated or treated with alkaline phosphatase (AP). **b**, Autoradiogram of ³⁵S-labelled wild-type and S51A mutant (51) eIF2 α that had been incubated with purified wild-type and mutant PERK. The phosphorylated (p-2 α) and non-phosphorylated (2 α) forms of eIF2 α were resolved by isoelectric focusing. **c**, Autoradiogram of ³⁵S-labelled luciferase translated from mRNA in reticulocyte lysates that had been pre-incubated with purified wild-type and mutant PERK for the indicated times. Haemoglobin staining serves as a loading control. **d**, Reduced synthesis rates of wild-type compared

with mutant PERK. Top, autoradiogram of ³⁵S-labelled PERK, immunoprecipitated from pulse-labelled COS-1 cells 48 h after transfection with the indicated PERK expression plasmids. Middle, northern blot of PERK and α -tubulin mRNA from identical plates, showing similar levels of wild-type and mutant PERK mRNA. Bottom, immunoblot showing similar levels of both PERK proteins and revealing that all the wild-type PERK in these cells had been activated. **e**, Interference with translation of a co-expressed green fluorescent protein (GFP) reporter gene by PERK. Cells transfected with a reporter plasmid expressing both GFP and wild-type or K618A mutant PERK were analysed by fluorescence-activated cell sorting for GFP activity (left panels), stained for PERK expression (top right four panels) and analysed for GFP reporter mRNA by northern blot and PERK protein by western blot (lower left panels).

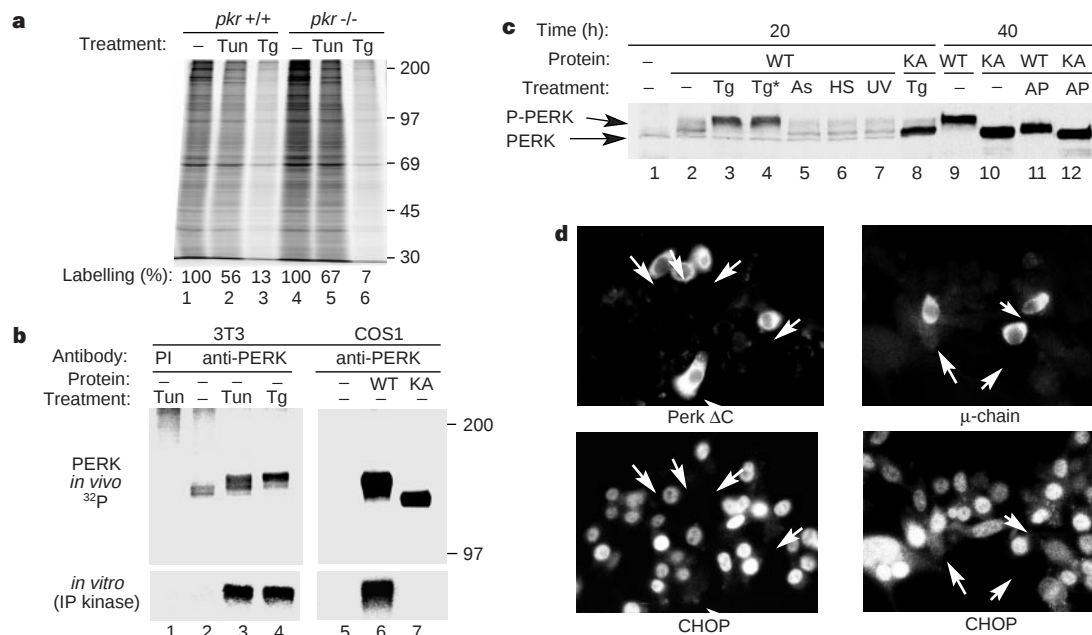


Figure 3 PERK is activated by ER stress. **a**, *pkr*^{-/-} cells are capable of attenuating translation in response to ER stress. Autoradiogram of total proteins synthesized in wild-type (+/+) and mutant (-/-) mouse embryo fibroblasts that had been untreated or exposed to tunicamycin (Tun) or thapsigargin (Tg) before a 30-min pulse label with ³⁵S-methionine. **b**, Top, autoradiogram of endogenous PERK immunoprecipitated from NIH3T3 cells (lanes 1-4) or of exogenous wild-type or K618A mutant PERK from COS-1 cells transfected with the indicated expression vectors (lanes 5-7). Cells had been preloaded with ³²P-labelled orthophosphate and exposed to a brief treatment with tunicamycin or thapsigargin as indicated. Lane 1 was immunoprecipitated with preimmune serum (PI). Bottom, *in vitro* autophosphorylation of PERK immunoprecipitated (IP) from cells treated as above. The left and right panels are exposures of the same gel that differ by a factor of 10 (left > right). **c**, Immunoblot of PERK in lysates of COS-1 cells

transfected with wild-type (lanes 2-7, 9, 11) or K618A mutant (lanes 8, 10, 12) PERK and collected 20 (lanes 1-8) or 40 (lanes 9-12) h after transfection. Lanes 3, 4 and 8 were treated with thapsigargin; lane 4 was pretreated with anisomycin (asterisk) to block protein synthesis. As, arsenite; HS, heat-shock; UV, ultraviolet irradiation. Extracts from lanes 11 and 12 were treated with alkaline phosphatase (AP) before being loaded on the gel. P-PERK, phosphorylated PERK on the gel. **d**, Immunostaining of CHOP in tunicamycin-treated NIH3T3 cells. Upper panels, photomicrographs of cells expressing the luminal domain of PERK (PERK Δ C) or the human immunoglobulin μ -chain, a protein that is retained in the ER and serves as a control for the effects of PERK Δ C. Lower panels, photomicrographs of the same cells stained with an antibody to CHOP. The arrows point to the cells expressing the exogenous PERK Δ C or μ -chain.

parallel yet distinct pathways²⁰. Our findings suggest considerable homology between the two processes: both are initiated by transmembrane ER-resident kinases, Ire1 or PERK, which have similar lumenal domains, and both kinases undergo ER-stress-induced activation, a step that may involve an activating autophosphorylation event^{21,22} (Fig. 3). Once activated, however, the two kinases appear to deliver distinct signals. On the basis of comparisons with yeast, mammalian Ire1 probably acquires endonuclease activity to cleave specific mRNA substrates⁹ and effect enhanced expression of upstream activators of the UPR²³. PERK, on the other hand, phosphorylates eIF2 α and attenuates translation. Thus, in mammalian cells two signalling pathways that are responsive to ER stress appear to have evolved by appending different effector functions to otherwise similarly structured upstream transducers.

Note added in proof: The rat orthologue of PERK has recently been identified as an eIF2 α kinase that is expressed at high levels in the pancreas, an organ that is heavily engaged in protein secretion and ER trafficking²⁶. This finding underscores the potential physiological significance of PERK's role in coupling folding to translation. □

Methods

Cloning of *Perk* cDNA. We used the insert from human EST clone IMAGE number 746909 as a hybridization probe to screen 3×10^6 clones from an unamplified NIH3T3 fibroblast cDNA library in λ -Zap.express (Stratagene). We identified 19 clones.

Expression plasmids. A C-terminal Myc tag and K618A mutation were incorporated into PERK by altering the *Perk* cDNA by polymerase chain reaction. The modified cDNAs were ligated into the pcDNA1/Amp expression vector (Invitrogen). PERK lacking the C terminus (PERK Δ C) was created by an internal *Sna*BI–*Sma*I deletion (removing amino acids 582–1,081 of the protein) of the Myc-tagged full-length cDNA. A bacterial expression plasmid encoding a GST-fusion protein with the cytoplasmic domain of PERK (amino acids 537–1,114) was constructed in pGEX 4T1 (Pharmacia). The purified fusion protein, of relative molecular mass $\sim 93,000$ ($M_r \sim 93K$), was used in the *in vitro* assays described below and to raise rabbit antisera.

Cell culture and transfection. COS-1, NIH3T3 and *pkr*^{+/+} and *pkr*^{−/−} mouse embryonic fibroblasts were maintained in DMEM supplemented with 10% fetal bovine serum. Unless otherwise indicated, cells were treated with $2 \mu\text{g ml}^{-1}$ tunicamycin for 4 h, $1 \mu\text{M}$ thapsigargin for 30 min, $100 \mu\text{M}$ sodium arsenite for 2 h, and heat-shock at 42°C for 2 h; pretreatment was with 27 J m^{-2} ultraviolet light or $10 \mu\text{g ml}^{-1}$ anisomycin for 45 min. COS-1 cells were transfected with $3 \mu\text{g}$ plasmid DNA per 60-cm dish by the DEAE-Dextran method. NIH3T3 cells were transfected with $1 \mu\text{g}$ DNA per 35-mm dish using Lipofectamine Plus (Life Technologies). Immunoblots were performed with rabbit anti-PERK serum diluted 1:10,000 nearly as described¹⁸ except that cell lysates were prepared in a HEPES–Triton–X100 buffer (see below) and contained $10 \mu\text{g}$ total protein. Immunostaining was done as described¹⁰.

Cell labelling and immunoprecipitation. Transfected COS-1 cells were starved for 30 min in methionine-free DMEM containing 10% dialysed fetal calf serum, then labelled for 2 h with $250 \mu\text{Ci ml}^{-1}$ ³⁵S-Translabel (ICN) in the same medium. Cells were lysed in RIPA buffer immediately after labelling (Fig. 2d) or after a 4 h cold-chase in complete media (Fig. 1d). The labelled protein was immunoprecipitated with the indicated antibodies and resolved by 6% SDS–PAGE. Where indicated, the immunopurified PERK was digested for 16 h at 37°C with 0.025 U endoglycosidase H or 0.25 U peptide N-glycosidase F according to the manufacturer's instructions (Boehringer). *In vivo* ³²P-labelling of PERK was performed by preloading 10-cm plates of NIH3T3 cells with $125 \mu\text{Ci ml}^{-1}$ ³²P-orthophosphate for 14 h before stress treatment and lysis in HEPES–Triton–X100 buffer (20 mM HEPES, pH 7.5, 150 mM NaCl, 1% Triton–X100, 10% glycerol, 1 mM EDTA, 10 mM sodium diphosphate, 100 mM NaF, 17.5 mM B-glycerophosphate, 1 mM phenylmethylsulphonyl fluoride, $4 \mu\text{g ml}^{-1}$ aprotinin, and $2 \mu\text{g ml}^{-1}$ pepstatin A). Nuclei were removed by brief centrifugation and the extracts were adjusted to a final concentration of 0.5% sodium deoxycholate and 0.1% SDS before immunoprecipitation with anti-PERK antiserum ($1 \mu\text{l}$).

In vitro phosphorylation. These reactions, containing immunopurified or

recombinant bacterially expressed wild-type or K618A mutant GST–PERK (50 ng), were done in $20 \mu\text{l}$ kinase buffer²⁴ at 30°C for 30 min with either $5 \mu\text{Ci}$ [γ -³²P]ATP ($4,500 \text{ Ci mmol}^{-1}$; ICN; Fig. 2a, 3b) or 0.1 mM unlabelled ATP (Fig. 2b). Where indicated, the reactions contained $\sim 10 \text{ ng}$ partially purified human eIF2 α , which was immunoprecipitated from the supernatant after incubation (Fig. 2a), or $0.5 \mu\text{l}$ *in vitro*-translated ³⁵S-methionine-labelled wild-type or mutant (S51A) eIF2 α (TNT reticulocyte lysate, Promega; Fig. 2b). The products in Fig. 2b were analysed by isoelectric focusing using Pharmalytes, pH 4–6.5, as described²⁵. To measure translational inhibition by PERK, glutathione–Sepharose beads ($1 \mu\text{l}$ bed volume) containing $\sim 50 \text{ ng}$ GST–PERK (wild-type or K618A) were pre-incubated for 1–15 min at room temperature with $12 \mu\text{l}$ of an *in vitro* translation mixture (Promega) before addition of $0.5 \mu\text{g}$ *in vitro*-transcribed capped luciferase mRNA. The labelled luciferase protein was resolved by 10% SDS–PAGE.

Received 5 October; accepted 16 November 1998.

- Prostko, C. R., Brostrom, M. A. & Brostrom, C. O. Reversible phosphorylation of eukaryotic initiation factor 2 alpha in response to endoplasmic reticular signaling. *Mol. Cell Biochem.* **127**–**128**, 255–265 (1993).
- Brostrom, C. O. & Brostrom, M. A. Regulation of translational initiation during cellular response to stress. *Prog. Nucleic Acid Res. Mol. Biol.* **58**, 79–125 (1998).
- Wong, W. L., Brostrom, M. A., Kuznetsov, G., Gmitter-Yellen, D. & Brostrom, C. O. Inhibition of protein synthesis and early protein processing by thapsigargin in cultured cells. *Biochem. J.* **289**, 71–79 (1993).
- Lee, A. Mammalian stress response: induction of the glucose-regulated protein family. *Curr. Biol.* **4**, 267–273 (1993).
- Kozutsumi, Y., Segal, M., Normington, K., Gething, M. J. & Sambrook, J. The presence of misfolded proteins in the endoplasmic reticulum signals the induction of glucose-regulated proteins. *Nature* **332**, 462–464 (1988).
- Gething, M. J. & Sambrook, J. Protein folding in the cell. *Nature* **355**, 33–45 (1992).
- Cox, J. S., Shamu, C. E. & Walter, P. Transcriptional induction of genes encoding endoplasmic reticulum resident proteins requires a transmembrane protein kinase. *Cell* **73**, 1197–1206 (1993).
- Mori, K., Ma, W., Gething, M. J. & Sambrook, J. A transmembrane protein with a cdc2+/CDC28-related kinase activity is required for signaling from the ER to the nucleus. *Cell* **74**, 743–756 (1993).
- Tirasophon, W., Welihinda, A. A. & Kaufman, R. J. A stress response pathway from the endoplasmic reticulum of the nucleus requires a novel bifunctional protein kinase/endoribonuclease (Ire1p) in mammalian cells. *Genes Dev.* **12**, 1812–1824 (1998).
- Wang, X.-Z. et al. Cloning of mammalian Ire1 reveals diversity in the ER stress responses. *EMBO J.* **19**, 5708–5717 (1998).
- Shamu, C. E. Splicing together the unfolded-protein response. *Curr. Biol.* **7**, R67–R70 (1997).
- Hinnebusch, A. G. The eIF-2 alpha kinases: regulators of protein synthesis in starvation and stress. *Semin. Cell Biol.* **5**, 417–426 (1994).
- Srivastava, S. P., Davies, M. V. & Kaufman, R. J. Calcium depletion from the endoplasmic reticulum activates the double-stranded RNA-dependent protein kinase (PKR) to inhibit protein synthesis. *J. Biol. Chem.* **270**, 16619–16624 (1995).
- Chen, J. J. & London, I. M. Regulation of protein synthesis by heme-regulated eIF-2 alpha kinase. *Trends Biochem. Sci.* **20**, 105–108 (1995).
- Berry, M. J., Knutson, G. S., Lasky, S. R., Munemitsu, S. M. & Samuel, C. E. Mechanism of interferon action. Purification and substrate specificities of the double-stranded RNA-dependent protein kinase from untreated and interferon-treated mouse fibroblasts. *J. Biol. Chem.* **260**, 11240–11247 (1985).
- Prostko, C. R., Dholakia, J. N., Brostrom, M. A. & Brostrom, C. O. Activation of the double-stranded RNA-regulated protein kinase by depletion of endoplasmic reticular calcium stores. *J. Biol. Chem.* **270**, 6211–6215 (1995).
- Fagard, R. & London, I. M. Relationship between phosphorylation and activity of heme-regulated eukaryotic initiation factor 2 alpha kinase. *Proc. Natl Acad. Sci. USA* **78**, 866–870 (1981).
- Wang, X.-Z. et al. Signals for the stressed endoplasmic reticulum induce C/EBP homologous protein (CHOP/GADD153). *Mol. Cell Biol.* **16**, 4273–4280 (1996).
- Price, B. & Calderwood, S. Gadd45 and Gadd153 messenger RNA levels are increased during hypoxia and after exposure of cells to agents which elevate the levels of glucose-regulated proteins. *Cancer Res.* **52**, 3814–3817 (1992).
- Brostrom, M. A., Prostko, C. R., Gmitter, D. & Brostrom, C. O. Independent signaling of grp78 gene transcription and phosphorylation of eukaryotic initiation factor 2 alpha by the stressed endoplasmic reticulum. *J. Biol. Chem.* **270**, 4127–4132 (1995).
- Shamu, C. E. & Walter, P. Oligomerization and phosphorylation of the Ire1p kinase during intracellular signaling from the endoplasmic reticulum to the nucleus. *EMBO J.* **15**, 3028–3039 (1996).
- Welihinda, A. A. & Kaufman, R. J. The unfolded protein response pathway in *Saccharomyces cerevisiae*. Oligomerization and trans-phosphorylation of Ire1p (Ern1p) are required for kinase activation. *J. Biol. Chem.* **271**, 18181–18187 (1996).
- Sidrauski, C. & Walter, P. The transmembrane kinase Ire1p is a site-specific endonuclease that initiates mRNA splicing in the unfolded protein response. *Cell* **90**, 1031–1039 (1997).
- Wu, S. & Kaufman, R. J. A model for the double-stranded RNA (dsRNA)-dependent dimerization and activation of the dsRNA-activated protein kinase PKR. *J. Biol. Chem.* **272**, 1291–1296 (1997).
- Savinova, O. & Jagus, R. Use of vertical slab isoelectric focusing and immunoblotting to evaluate steady-state phosphorylation of eIF2 alpha in cultured cells. *Methods* **11**, 419–425 (1997).
- Shi, Y. et al. Identification and characterization of pancreatic eukaryotic initiation factor 2 alpha-subunit kinase, PEK, involved in translational control. *Mol. Cell Biol.* **18**, 7499–7509 (1998).

Acknowledgements. We thank R. Schneider for the eIF2 α expression plasmids, antisera and purified protein; M. Green for the immunoglobulin μ -chain expression plasmid; B. Williams for the *pkr*^{−/−} cells; G. Kreibich for the antisera against ribophorin; and X.-Z. Wang for the cDNA library. This work was supported by NIH grants and an NIH training grant (to H.P.H.). D.R. is a Stephen Birnbaum Scholar of the Leukemia Society of America.

Correspondence and requests for materials should be addressed to D.R. (e-mail: ron@saturn.med.nyu.edu). The nucleotide sequence of *perk* has been submitted to GeneBank under the accession number AF076681.

Controlling a chain reaction

A selection of new products aimed at the DNA amplification market features PCR kits, and the paraphernalia needed to make use of them.

PCR Cloner kit

From Flowgen www.flowgen.co.uk

Designed to speed-up handling of PCR products

The Prime PCR Cloner kit from Flowgen is claimed to be 4–16 times faster than conventional T-vector kits. Using the kit it should be possible to purify, clone and transform any PCR product within 90 minutes and to obtain identifiable colonies within 12–14 hours. A novel ligation reagent is used in order to avoid the need for overnight ligation. The kit comes as a complete system with PCR purification spin columns and control reagents, in two forms with different sizes of cut-off. PCR Cloner is designed to be compatible with both blunt-ended fragments and those with T-A overhangs.

Reader Service No. 100

GeneRuler DNA Ladder

From Fermentas www.fermentas.com

Markers from 50 to 1,000 base pairs, in 50- and 100-base-pair steps

The GeneRuler 50 bp DNA Ladder yields the following 13 discrete fragments (in base pairs): 1,000, 900, 800, 700, 600, 500, 400, 300, 250, 200, 150, 100, 50. In the production process, pE-50 DNA is completely digested by *Eco*147I and *Pvu*II, phenol extracted, ethanol precipitated and dissolved in 10 mM Tris-HCl (pH 7.6), 1 mM EDTA. The 'ruler' is supplied with 6 × loading dye solution (0.2% bromophenol blue, 0.2% xylene cyanol, 60% glycerol, 60 mM EDTA) and at a DNA concentration of 0.5 mg per ml. The bright bands at 500 bp, 200 bp and 100 bp serve as the reference.

Reader Service No. 101

pGEM-T vectors

From Promega UK www.euro.promega.com/uk

Kits for the isolation and cloning of PCR products

One hour is the time claimed for ligation of PCR products using the new pGEM-T vector systems. A single 3' overhanging thymidine at each end of the vector is designed to improve the efficiency of ligation to PCR products made by most non-proof-reading thermostable polymerases. The Promega range of pGEM-T and pGEM-T Easy Vector Systems has now been re-configured to contain a new 2 × rapid ligation buffer, reducing the time needed for the ligation reaction from overnight to 1 hour.

Reader Service No. 102



Say goodbye to overnight ligation: purify, clone and transform PCR products with change from 2 hours.

Pfu DNA polymerase

From Promega www.promega.com

A high-fidelity PCR enzyme tested for activity and contaminating nucleases

Pfu DNA polymerase is recommended for PCR applications requiring high fidelity — including cloning, DNA expression and mutation analysis. Native *Pfu* DNA polymerase is now available from Promega in two sizes, 100 and 500 µl, both including 10 × reaction buffer containing MgSO₄.

Reader Service No. 103

TaKaRa Z-Taq DNA polymerase

From Takara Shuzo

A polymerase developed with speed in mind

TaKaRa Z-Taq is a novel DNA polymerase developed with a view to rapid PCR. *TaKaRa Z-Taq* is suitable for high-throughput multi-sample processing on a single cyclor. In handling 10,000 samples of 1 kbp, for example, the enzyme completes amplification of 1 kbp in approximately 20 minutes compared to 1.5 hours with a conventional *Taq*. All the necessary PCR reactions should therefore be completed in about 1.5 days compared to 6.5 days with a general *Taq*.

Reader Service No. 104

REDTaq DNA polymerase

From Sigma www.sigma-aldrich.com

A new rendition of Taq DNA polymerase with a dye to make it more visible

Sigma's new enzyme, REDTaq, delivers the same performance as Sigma's standard *Taq*, but it is easier to use. Addition of REDTaq to a tube produces a thin red layer that is visible at the bottom. This colour coding eliminates the uncertainty that can occur with interruptions in pipetting: when the red colour is

uniform, the solution is thoroughly mixed. REDTaq requires no loading buffers since the post-PCR samples are dense enough to be loaded directly onto an agarose gel. The red dye works as a tracking dye, migrating just faster than bromophenol blue. REDTaq is formulated at 1 unit per µl to help with accurate pipetting.

Reader Service No. 105

HotStarTaq

from Qiagen www.qiagen.com

A form of Taq DNA polymerase developed for hot-start PCR

HotStarTaq DNA polymerase is a modified form of Qiagen's *Taq* DNA polymerase designed to minimize nonspecific amplification, primer-dimer formation, and background, maximizing yield of the specific PCR product. The enzyme is active only after a 15-minute, 95 °C incubation, preventing extension of primers that anneal nonspecifically at low temperatures during reaction setup and the initial thermal-cycler heating phase. Thus all reaction components can be assembled at room temperature.

Reader Service No. 106

QuanToxBlunt cloning vector

From Quantum Biotechnologies
www.quantumbiotech.com

Positive discrimination, claimed to give efficient cloning and a low background

The introduction of new thermostable polymerases that produce blunt-ended PCR products has created the need for new PCR cloning methods. Rather than using rare restriction enzymes, special PCR strategies and/or expensive antibiotic selection, the GATA-1 gene present in the QuanToxBlunt vector codes for a toxic peptide that inhibits

the growth of bacterial cells when the expression is induced. Ligation of DNA fragments into the blunt cloning site disrupts the gene coding for the toxic peptide. Recombinant molecules are then non-toxic while vector molecules re-ligated on themselves are toxic.

Reader Service No. 107

PCR Master Mix

From Marsh Biomedical www.biomar.com

A ready-made mixture to reduce preparation time for PCR reactions

The Abgene PCR Master Mix is designed to increase the speed and accuracy of high-throughput PCR. Each mix contains *Taq* polymerase, dNTPs, buffer and $MgCl_2$. Just add template and primers to complete the reaction. The mixes are available with varying concentrations of $MgCl_2$ to suit any PCR protocol in vials containing 1.8 ml of 1.1 × Master Mix, or pre-aliquoted into individual tubes, strip tubes or 96-well plates. Suppliers are Marsh Biomedical, Rochester, NY, who can be called on (800) 445-2812.

Reader Service No. 108

Competitive RT-PCR

From Ambion www.ambion.com

Kits to simplify mRNA quantitation

Ambion's Competitive Quantitative RT-PCR kits are designed for the competitive quantitation of mRNA using a synthetic exogenous RNA standard added to RT-PCR reactions. This RNase-proof standard acts as a control for variations in RNA isolation, reverse transcription, and amplification. Twelve kits are available, each containing a gene-specific RNA competitor and primers to amplify both the competitive and endogenous RNA species. Call (800) 888-8804 in the US.

Reader Service No. 109

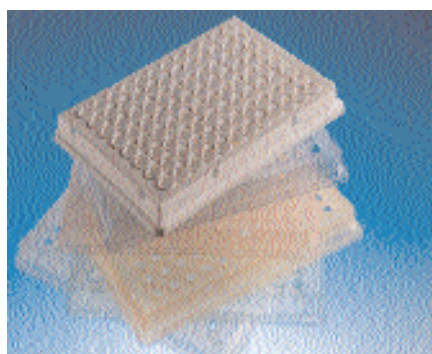
Skirted Thermo-Fast 96

From Advanced Biotechnologies

www.adbio.co.uk

Raised rims in 96-well plate form

The latest addition to the Advanced Biotechnologies Thermo-Fast PCR plate range is the Skirted Thermo-Fast 96 plate, an alternative to existing polycarbonate plates compatible with most 96-well thermal cyclers and robotic workstations. The new plates are injection-moulded in polypropylene. Injection moulding is claimed to give less well-to-well variation than the vacuum-forming used for polycarbonate plates. The raised rims around each tube are to ensure that contact is made with the chosen sealing medium and to reduce the problems associated with evaporation, thus giving the option of performing oil-free PCR. DNase- and RNase-free, the Skirted Thermo-Fast 96 comes in natural, blue, green, red and yellow.



Spot the difference: it's the rims that distinguish these 96 wells from the conventional kind.

For fluorescent or chemiluminescent studies, the plate comes in opaque black or white.

Reader Service No. 110

Calypso RT-PCR system

From Tetra Link www.tetra-link.com

A single-tube RT-PCR system using AMV reverse transcriptase

Calypso is a single-tube RT-PCR system combining AMV reverse transcriptase with Tetra Link's Accurase polymerase mix. The system offers sensitive, high-fidelity, reproducible RNA analysis. Tetra Link claim that, compared to RT-PCR systems using *Taq*, *Tfl* or *Tth* DNA polymerases, the combination of the three-enzyme Accurase mix and AMV RT allows amplification of longer products coupled with improved product yield, a threefold increase in fidelity and increased sensitivity due to the efficiency of the three-enzyme system. The system requires between 10 pg and 1 µg of total RNA or between 1 pg and 100 ng of poly (A)⁺ RNA to obtain a visible band.

Reader Service No. 111

TrueSprinter RT-PCR kit

From Hybaid www.hybaid.co.uk

A kit containing the reagents for RT-PCR, plus λ-DNA and primers

The TrueSprinter kit is based on TrueScript reverse transcriptase (MMLV) which lacks RNase H activity and allows synthesis of cDNA molecules of up to 20 kb. cDNA products are amplified with the ProofSprinter mixture of *Taq* and *Pwo*, included in the kit, for high yield of full-length PCR products.

Reader Service No. 112

TOPO XL kit

from Invitrogen www.invitrogen.com

A PCR cloning kit designed for use with large PCR fragments

The TOPO XL PCR cloning kit is designed specifically for high-efficiency cloning of long (3–10 kb) PCR products. The kit uses a topoisomerase I-activated positive selection

vector, pCR-XL-TOPO, to achieve 5-minute efficient cloning and low background. With a 15-minute gel purification step 77–80% recombination can be obtained. The kit is claimed to achieve 80% recombinant for cloning a 7-kb PCR fragment.

Reader Service No. 113

RT-PCR primers

from Continental Lab Products

www.conlab.com

A new range of RT-PCR primers

Designed specifically for PCR conducted from mRNA sources, Continental's range of RT-PCR primers are selected for efficient generation of PCR products across intron–exon boundaries, to ensure that the signal is derived from expressed genes. All sets are pre-tested on cDNA from cell lines or cultures known to express the appropriate analyte. The ready-to-use aliquots contain 2.5 nanomoles of each primer.

Reader Service No. 114

Dynabeads mRNA DIRECT

From Dynal www.dynal.no

mRNA isolation for RT-PCR amplification

The Dynabeads mRNA DIRECT Micro kit is a new microscale kit for isolation of mRNA directly from cells, plant and animal tissue for RT-PCR amplification. The kit combines poly(A)⁺ RNA isolation and RT-PCR and is designed to isolate highly purified and intact mRNA from small numbers of cells and small tissue samples. The isolated mRNA can be used directly in RT-PCR amplification without eluting the mRNA from the beads. The direct isolation is performed in 15 minutes without having to prepare total RNA or perform any other purification steps. The oligo(dT)₂₅ bound to the Dynabeads surface is used both to capture the mRNA and as a primer for the reverse transcription of first-strand cDNA. Dynabeads solid-phase cDNA synthesis technology is compatible with various commercially available cDNA synthesis kits.

Reader Service No. 115

PCR-ready DNA

From Epicentre Technologies

www.epicentre.com

A kit to prepare PCR-ready DNA from any source in under an hour

The Epicentre MasterPure kits come in three versions, for purification of DNA, RNA or both DNA and RNA. The kits can perform 200, 100 and 200 purifications respectively. All manipulations can be performed in 1.5-ml microcentrifuge tubes. The recovered nucleic acids are claimed to be highly pure, with the ratio of absorbance at 260 and 280 in the range 1.8–2.0.

Reader Service No. 116

Multiplex PCR kit

From Maxim Biotech www.maximbio.com

A PCR kit targeted at high-throughput applications

Multiplex PCR (MPCR) kits utilize a method of detecting multiple gene expression by the amplification of many genes under identical conditions in a single reaction tube. This technique offers an alternative to the traditional time-consuming northern blot and RNase protection assays (RPA) for determining the abundance of a specific mRNA in a total or poly(A) RNA sample. Multiplex PCR is offered as an economical, high-throughput method of monitoring gene expression. In the UK, the Maxim range is available from Bio-Stat Diagnostics.

Reader Service No. 117

PCR Clean-Up kit

From Mo Bio www.mobio.com

A PCR Clean-Up kit to eliminate the need for an agarose gel

The purpose of the UltraClean PCR Clean-Up kit is to allow researchers to purify PCR products directly from a PCR reaction without running an agarose gel first. A silica spin filter is used to selectively bind the PCR product, and unwanted reaction components are passed through the filter. The PCR product is then eluted from the spin filter into certified DNA-free water (supplied) or TE. This simple procedure takes only 3 minutes.

Reader Service No. 118

Quadruple-density microplates

From Porvair Sciences
www.porvair-science.com

Compact 384 deep-well plates for high-throughput screening assays

A quadruple-density microplate with double the sample capacity of standard plates has been added to the Porvair Sciences microplate range. The maximum capacity of each of its 384 wells is over 120 µl, but the system retains the same footprint as a 96-well plate, making it compatible with most scanners, analysers, liquid handling and automated sample processors. This makes the new plate suitable for sample transfer, high-throughput screening, most chemical assays and many other common applications in the pharmaceutical and life science markets. The polystyrene plate comes in three different forms clear, opaque white and solid black. The former is intended for luminescence- and scintillation-based assays, where maximum light reflection is required, while the latter provide the all-absorbing background needed for fluorescence techniques.

Reader Service No. 119

Apex PCR Certified Water

From Apex

Water, but water specially prepared for PCR applications

Produced specifically as a component for PCR reactions, Apex PCR-certified water is tested for PCR compatibility and certified to be free of all RNase, DNase, and DNA contamination. The water is useful for highly sensitive PCR reactions and comes in 4 × 1.25 ml and 1 × 8.0 ml package configurations, double-pouched for added protection.

Reader Service No. 120

PCR Express gradient block

From Hybaid www.hybaid.com

A new approach to gradient cycling

Hybaid has recently launched the new PCR Express gradient thermal cycler. The gradient block is interchangeable and complements the existing range of PCR Express blocks. The block is available in 0.2 ml or 0.5 ml formats and allows a temperature gradient of up to 15 °C to be set. Hybaid has introduced new software and 'Active Tube' temperature control with a view to increasing temperature and uniformity. Active Tube Control measures actual in-sample temperatures, ensuring samples achieve programmed temperatures and times.

Reader Service No. 121

GENESIS Workstation

From Tecan www.tecan.ch

A workstation suited for use in PCR

Tecan's GENESIS Workstation can be customized for molecular biology applications such as PCR, DNA extraction or sequencing reaction preparation. A new, low volume option allows users to pipette reliably to 0.5 µl, especially useful for PCR reactions where exotic reagents and DNA samples are used. The RoMa (robotic manipulator) arm



Let there be blue light: the GENESIS Workstation tackles volumes down to 0.5 µl.

moves microplates to systems such as magnetic holders, thermal cyclers or microplate storage hotels placed on or around the worktable. The GENESIS Workstation runs under a Windows NT-based operating system called Flexible Assay Composer and Task Scheduler (FACTS). This software allows users to run multiple batches of samples, because individual modules can be programmed to work simultaneously.

Reader Service No. 122

Tgradient thermocycler

From Biometra www.biometra.de

A modular-block gradient thermocycler optimized for DNA amplification

Biometra's new thermocycler features a very wide temperature gradient of up to 40 °C temperature difference, high block uniformity courtesy of a gold-plated silver block, and fast temperature cycles. Ease of use is billed as an important advantage of the system as a result of its modular construction, adjustable heated lid with defined pressure, big display and intuitive programming to provide flexibility.

Reader Service No. 123

PCR poster

From Eppendorf www.eppendorf.com

A poster for the lab wall with tips for PCR

The latest edition of Eppendorf's publication *BioNews* (Issue 11) includes a PCR poster created to give newcomers to PCR an overview of the common rules, and an indication of the concentrations that should be used when designing experimental set-ups. *BioNews* is available free from Eppendorf. The company also invites users and readers to send suggestions for changes, or new features for the poster that might increase its usefulness.

Reader Service No. 124

PCR plastics

From Robbins Scientific www.robsci.com

A new brochure covers a range of PCR plastics products

An electronic copy of the new brochure for the Robbins Scientific range of disposable products for PCR is available from the company's website (see above), or via e-mail from custserv@robsci.com. A 4-colour paper version is available by post. Among the products described is the CyclePlate-ET PCR plate, a new range of thin-wall PCR plates designed for oil-free sample processing in high-performance thermal cycling.

Reader Service No. 125

These notes are compiled in the Nature office from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.